Implementasi SMOTE Untuk Mengatasi *Imbalance Class* Pada Berita Online Menggunakan Metode *K-Nearest Neighbor* (KNN)

1st Andreas Sitanggang
Fakultas Teknik Elektro
Telkom University Purwokerto
Purwokerto, Indonesia
@ittelkom-pwt.ac.id

2nd M Shinta Rhomandhona,ST., M.T.
Fakultas Teknik Elektro
Telkom University Purwokerto
Purwokerto, Indonesia
@ittelkom-pwt.ac.id

3rd Mas Aly Afandi, S.ST., M.T.
Fakultas Teknik Elektro
Telkom University Purwokerto
Purwokerto, Indonesia
@ittelkom-pwt.ac.id

***Abstrak* — Berdasarkan survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), jumlah pengguna internet di Indonesia mencapai 215,63 juta orang pada periode 2022-2023, meningkat sebesar 2,67% dibandingkan periode sebelumnya. Peningkatan ini menyebabkan lonjakan jumlah berita online yang memerlukan pengelolaan data yang lebih baik, terutama dalam menangani ketidakseimbangan kelas data set pada klasifikasi data. Penelitian ini bertujuan untuk mengatasi masalah tersebut dengan menerapkan teknik SMOTE, yang menghasilkan sampel baru untuk kelas data set minoritas guna meningkatkan representasi data. Selain itu, algoritma KNN digunakan untuk mengevaluasi pengaruh kombinasi SMOTE dan KNN terhadap performa model klasifikasi. Evaluasi dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-Score. Hasil penelitian menunjukkan bahwa penerapan SMOTE berhasil meningkatkan performa model klasifikasi. Kombinasi terbaik diperoleh pada nilai parameter k=1, dengan akurasi sebesar 62,50%, presisi 58,39%, recall 86,96%, dan F1-Score 69,87%. Dibandingkan dengan model sebelum penerapan SMOTE, terjadi peningkatan performa akurasi dari 58,33%, presisi dari 49,56%, dan F1-Score dari 63,28%, sambil mempertahankan recall 87,50%. Penelitian ini membuktikan bahwa SMOTE efektif dalam menangani ketidakseimbangan kelas data set, menghasilkan prediksi model yang lebih akurat dan seimbang. Hasil penelitian memberikan kontribusi dalam pengelolaan data berita online untuk mendukung kualitas klasifikasi yang lebih baik.**

***Kata kunci*— AI, KNN, SMOTE, Berita Online**

I. PENDAHULUAN

Berdasarkan hasil survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), pengguna internet di Indonesia mencapai 215,63 juta orang pada periode 2022-2023. Angka ini menunjukkan peningkatan sebesar 2,67% dibandingkan periode sebelumnya yang mencatatkan 210,03 juta pengguna. Jumlah tersebut setara dengan 78,19% dari total penduduk Indonesia yang mencapai 275,77 juta jiwa, meningkat sebesar 1,17% dari tingkat penetrasi internet pada 2021-2022 yang tercatat sebesar 77,02%. Berdasarkan analisis lebih lanjut, tingkat penetrasi internet menunjukkan perbedaan berdasarkan jenis kelamin. Pengguna internet laki-laki mencapai tingkat penetrasi 79,32%, sedikit lebih tinggi dibandingkan dengan perempuan yang berada pada angka 77,36%. Selain itu, perbedaan penetrasi juga terlihat berdasarkan lokasi geografis. Di kawasan perkotaan, tingkat penetrasi internet pada periode 2022-2023 tercatat sebesar 77,36%, sedangkan di kawasan perdesaan tingkat penetrasi bahkan lebih tinggi, yaitu 79,79%. Data ini mencerminkan dinamika penggunaan internet di Indonesia yang terus berkembang, baik secara keseluruhan maupun dalam konteks demografis tertentu, menunjukkan potensi besar bagi pengembangan teknologi digital dan infrastruktur komunikasi di berbagai wilayah [1].

Berita merupakan sumber informasi penting yang menyajikan peristiwa terkini dan dapat ditemukan di berbagai media massa seperti surat kabar, televisi, serta platform lainnya. Dalam perkembangannya, kemajuan teknologi informasi telah secara signifikan mempermudah proses penyebaran berita melalui *media online*. *Media online* adalah jenis media massa yang menggunakan internet sebagai basisnya, memungkinkan akses informasi yang nyaman dan *real-time*. Pembaca kini tidak lagi perlu menunggu hingga keesokan hari untuk mendapatkan kabar terbaru; cukup dengan memanfaatkan internet, berita dapat diakses dengan cepat kapan saja dan di mana saja.

Namun, meskipun kemudahan akses ini sangat membantu, banyaknya informasi yang tersedia secara daring belum sepenuhnya diimbangi dengan kualitas dan objektivitas dari informasi itu sendiri. Situasi ini menimbulkan tantangan baru bagi pembaca, yaitu memilah dan menyaring informasi yang relevan serta dapat dipercaya di tengah arus data yang melimpah. Proses ini tidak hanya memakan waktu tetapi juga mengharuskan pembaca memiliki kemampuan literasi informasi yang baik. Oleh karena itu, diperlukan sistem yang mampu secara efektif memilah atau mengklasifikasikan informasi berdasarkan tingkat objektivitas dan relevansinya, sehingga memudahkan pembaca untuk memperoleh informasi yang berkualitas [2].

Di era digital, kecepatan dan akurasi informasi menjadi kunci utama dalam mendukung pengambilan keputusan, baik secara individu maupun kolektif. Media online kini telah menjadi sumber utama bagi masyarakat untuk mendapatkan berita terkini. Namun, muncul tantangan signifikan dalam pengelolaan dan klasifikasi berita daring, terutama terkait ketidakseimbangan jumlah berita di antara berbagai kategori. Kondisi ini dapat menyebabkan model klasifikasi yang digunakan dalam pengolahan berita cenderung mengabaikan kategori atau kelas berita yang jumlahnya lebih sedikit (kelas minoritas). Akibatnya, hasil prediksi dari model tersebut menjadi kurang akurat dan tidak seimbang, yang pada akhirnya mengurangi kepercayaan terhadap sistem klasifikasi itu sendiri [3]. Untuk mengatasi permasalahan ini, dibutuhkan pendekatan khusus yang mampu mengatasi ketidakseimbangan kelas dalam berita daring. Pendekatan tersebut harus mampu mempertimbangkan keberagaman kategori berita secara proporsional, sehingga dapat menghasilkan model klasifikasi yang lebih akurat, andal, dan memberikan manfaat optimal bagi pengguna.

Untuk mengatasi ketidakseimbangan kelas dalam klasifikasi informasi online, Penulis menggunakan *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE bekerja dengan menghasilkan sampel sintetik dari kelas minoritas, sehingga dapat meningkatkan keterwakilan dan kemampuan prediksi model untuk kelas minoritas [4]. Selain itu, Penulis juga menggunakan algoritma *K-Nearest Neighbors* (KNN). KNN menggunakan informasi dari tetangga terdekat untuk membuat prediksi, sehingga memungkinkannya berkontribusi aktif dalam pengelolaan ketidakseimbangan kelas. Penggabungan antara SMOTE dan KNN dapat menciptakan model klasifikasi berita online yang lebih andal, mampu membuat prediksi berimbang di seluruh kategori berita.

II. KAJIAN TEORI

A. Penelitian Terkait

Penelitian pertama pada tahun 2020 mengenai SMOTE dimana pada penelitian ini mengintegrasikan algoritma *Naïve Bayes* dengan teknik SMOTE untuk menangani ketidakseimbangan data dalam proses klasifikasi. Hasil penelitian menunjukkan bahwa penggunaan *Naïve Bayes* yang dipadukan dengan SMOTE berhasil meningkatkan performa klasifikasi secara signifikan. Sebanyak 1.131 data berhasil diklasifikasikan dengan benar, sementara 72 data tidak diklasifikasikan dengan benar. Sebaliknya, tanpa menggunakan SMOTE, hanya 818 data yang berhasil diklasifikasikan dengan benar, sementara 60 data tidak diklasifikasikan dengan benar. Hasil ini menunjukkan bahwa penerapan SMOTE memberikan kontribusi yang nyata dalam meningkatkan akurasi model, meskipun terdapat sedikit peningkatan pada data yang tidak terklasifikasi dengan baik [5].

Penelitian kedua dilakukan pada tahun 2021, yang memfokuskan analisisnya pada *rating* aplikasi Shopee menggunakan metode *Decision Tree* berbasis SMOTE. Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* tanpa menggunakan SMOTE menghasilkan nilai presisi sebesar 99,89%, *recall* sebesar 99,88%, dan AUC (*Area Under Curve*) sebesar 0,950. Ketika SMOTE diterapkan, nilai presisi meningkat menjadi 99,98%, sementara AUC meningkat menjadi 0,999. Namun, nilai *recall* tidak mengalami perubahan signifikan. Selisih nilai akurasi hanya sebesar 0,02% dan peningkatan AUC sebesar 0,049. Hal ini menunjukkan bahwa penerapan SMOTE memberikan pengaruh signifikan terhadap metrik tertentu seperti AUC, meskipun dampaknya terhadap metrik lainnya seperti presisi dan *recall* cenderung minimal [6].

Penelitian selanjutnya pada tahun 2021 membandingkan performa algoritma *Random Forest* dan *XGBoost* dalam konteks prediksi cuaca menggunakan SMOTE. Hasil penelitian menunjukkan bahwa *Random Forest* dengan SMOTE mencapai akurasi tertinggi sebesar 95,59%, sementara *XGBoost* menghasilkan akurasi tertinggi sebesar 94,34% tanpa menggunakan SMOTE. Penelitian ini mengungkapkan bahwa efektivitas SMOTE dapat bervariasi tergantung pada algoritma yang digunakan, di mana *Random Forest* lebih efektif dalam memanfaatkan SMOTE untuk meningkatkan performa klasifikasi dibandingkan *XGBoost* [7].

Penelitian lainnya memfokuskan pada penerapan SMOTE-NCL untuk mengatasi ketidakseimbangan kelas dalam klasifikasi penyakit jantung koroner menggunakan *Support Vector Machine (SVM)*. Hasil penelitian menunjukkan bahwa SVM tanpa menggunakan balancing data menghasilkan akurasi sebesar 76,60%. Ketika SMOTE diterapkan, akurasi meningkat menjadi 80,85%. Kombinasi SMOTE-NCL dengan SVM memberikan hasil terbaik dengan akurasi mencapai 85,10%, yang menunjukkan bahwa SMOTE-NCL lebih efektif dibandingkan SMOTE standar dalam meningkatkan performa klasifikasi pada *Dataset* yang tidak seimbang [8].

Penelitian selanjutnya menyoroti penerapan SMOTE untuk klasifikasi *performance rating* iklan televisi menggunakan algoritma *Artificial Neural Network (ANN)*. Hasil penelitian menunjukkan bahwa model ANN tanpa SMOTE memiliki akurasi sebesar 86,35%. Setelah menerapkan SMOTE, akurasi meningkat menjadi 87,06%. Hal ini menunjukkan bahwa SMOTE dapat meningkatkan performa klasifikasi meskipun peningkatannya relatif kecil. Penelitian ini menegaskan pentingnya penerapan teknik balancing data untuk meningkatkan kualitas hasil klasifikasi dalam konteks data yang tidak seimbang [4].

B. SMOTE

SMOTE (*Synthetic Minority Over-sampling Technique*) adalah salah satu metode yang digunakan untuk menangani ketidakseimbangan kelas dalam *Dataset*. Teknik ini bekerja dengan cara menghasilkan data sintetis untuk kelas minoritas alih-alih hanya menduplikasi data yang sudah ada. SMOTE menggunakan pendekatan berbasis interpolasi, di mana data sintetis dibuat dengan memilih data dari kelas minoritas secara acak dan kemudian menggabungkannya dengan data tetangga terdekat (berdasarkan jarak Euclidean) dalam ruang fitur. Data baru ini terletak di antara titik-titik data asli, sehingga menghasilkan variasi yang lebih besar dalam kelas minoritas dan membantu model untuk mempelajari pola yang lebih representatif.

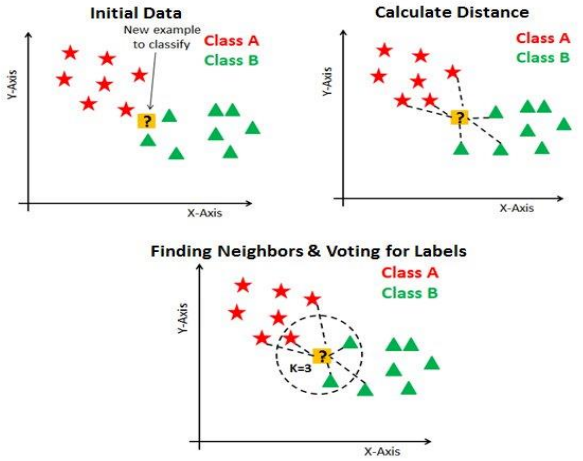
Proses SMOTE memiliki beberapa langkah utama. Pertama, data kelas minoritas diidentifikasi, dan untuk setiap titik data, sejumlah tetangga terdekat dihitung menggunakan algoritma seperti KNN Kedua, satu atau lebih tetangga terdekat dipilih secara acak, dan titik data sintetis baru dibuat dengan interpolasi linear antara data asli dan tetangganya.

Dengan pendekatan ini, SMOTE tidak hanya meningkatkan jumlah data kelas minoritas tetapi juga mendistribusikan data tambahan tersebut dalam ruang fitur, sehingga mengurangi kemungkinan overfitting yang sering terjadi pada metode duplikasi sederhana. Langkah-langkah penerapan SMOTE dimulai dengan menghitung jarak antar data pada data minoritas, kemudian menentukan nilai persentase SMOTE kemudian menentukan bilangan k terdekat dan terakhir membuat data sintetik.

C. K -Nearest Neighbour (KNN)

Algoritmek tetangga terdekat (bahasa Inggris: *k-nearest neighbour algorithm*, disingkat k-NN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pemelajaran yang jaraknya paling dekat dengan objek tersebut. Data pemelajaran digambarkan ke ruang berdimensi banyak dengan tiap-tiap dimensi mewakili tiap ciri/fitur dari data. Klasifikasi data baru dilakukan dengan mencari label ktetangga terdekat. Label terbanyak yang muncul menjadi label data baru. Bila k=1, data baru dilabeli dengan label tetangga terdekat [24].

Untuk menghitung kedekatan antar data, KNN sering menggunakan metrik jarak, seperti jarak Euclidean, Manhattan, atau Minkowski. Algoritma ini bersifat instance-based, yang berarti tidak ada model eksplisit yang dibangun selama proses pelatihan; data yang ada digunakan langsung untuk membuat prediksi saat ada data baru. Kelebihan dari KNN adalah kesederhanaannya dan tidak memerlukan pelatihan (model bebas pelatihan), serta kemampuannya untuk menangani data dengan berbagai tipe atribut. Namun, KNN juga memiliki kelemahan, yaitu kinerjanya bisa menurun dengan sangat signifikan jika jumlah data sangat besar atau jika data memiliki banyak dimensi (fitur), karena perhitungan jarak akan semakin mahal secara komputasi.

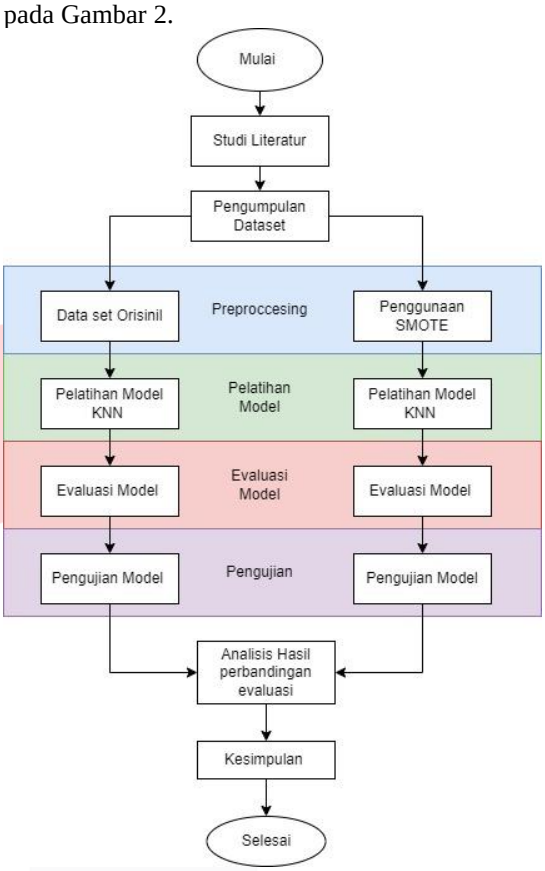


GAMBAR 1
Ilustrasi KNN

Pada Gambar 1, menjelaskan algoritma KNN untuk klasifikasi data. Data awal terdiri dari dua kelas: *Class* A (bintang merah) dan *Class* B (segitiga hijau), dengan data baru (tanda tanya kuning) yang belum diketahui kelasnya. Algoritma menghitung jarak antara data baru dengan semua titik data yang ada menggunakan metrik tertentu, seperti *Euclidean distance*, untuk menemukan tetangga terdekat. Dengan nilai k = 3, tiga tetangga terdekat terdiri dari dua anggota *Class* A dan satu anggota *Class* B. Berdasarkan prinsip voting mayoritas, data baru diklasifikasikan ke dalam *Class* A karena dua dari tiga tetangga terdekat berasal dari kelas tersebut [25].

III. METODE PENELITIAN

Penelitian ini dilakukan dalam beberapa tahap yang dapat dilihat pada Gambar 2.



GAMBAR 2
Alur Penelitian

Berdasarkan Gambar 2, penelitian dimulai dengan studi literatur. Pada tahap ini, peneliti mempelajari berbagai literatur yang relevan untuk memahami konsep ketidakseimbangan kelas, penerapan SMOTE, dan algoritma KNN. Studi ini bertujuan untuk mengidentifikasi metode dan pendekatan terbaik yang akan diterapkan dalam penelitian. Peneliti juga mengevaluasi teori-teori yang mendukung penggunaan SMOTE dan parameter yang optimal untuk algoritma KNN, seperti nilai kk pada *K-Nearest Neighbors*.

Tahap berikutnya adalah pengumpulan *Dataset* yang menjadi dasar untuk membangun model. *Dataset* yang digunakan harus mencerminkan masalah ketidakseimbangan kelas, di mana salah satu kelas memiliki jumlah data yang jauh lebih besar dibandingkan kelas lainnya. Peneliti memastikan bahwa *Dataset* yang digunakan relevan dengan konteks penelitian, seperti data komentar berita yang memerlukan klasifikasi. Setelah *Dataset* terkumpul, dilakukan pra-pemrosesan data, termasuk pembersihan data, normalisasi, dan pembagian *Dataset* menjadi dua jalur: *Dataset* orisinal tanpa modifikasi dan *Dataset* yang telah diolah menggunakan teknik SMOTE. Jalur SMOTE bertujuan untuk menyeimbangkan data kelas minoritas dengan menciptakan data sintetis yang menyerupai pola asli.

Pada tahap pembangunan model, algoritma KNN diterapkan pada kedua jalur *Dataset* tersebut. Model pertama menggunakan *Dataset* orisinal untuk mendapatkan baseline

performa, sementara model kedua menggunakan *Dataset* hasil SMOTE untuk membandingkan peningkatan kinerja. Peneliti juga mengeksplorasi parameter k dengan mencoba berbagai nilai dari k = 1 hingga k=25 untuk menentukan konfigurasi terbaik yang menghasilkan akurasi tertinggi. Setelah model dibangun, dilakukan evaluasi model menggunakan metrik seperti akurasi, presisi , *recall*, dan *F1-Score* untuk menilai performa model dalam mendeteksi kelas minoritas.

Setelah evaluasi, tahap berikutnya adalah analisis hasil perbandingan antara model yang dilatih dengan *Dataset* orisinal dan *Dataset* hasil SMOTE. Peneliti mengevaluasi dampak SMOTE terhadap kemampuan algoritma KNN dalam mengenali pola dari kelas minoritas, serta apakah teknik ini mampu mengurangi bias terhadap kelas mayoritas. Jika hasilnya memuaskan, kesimpulan diambil berdasarkan temuan, yang mencakup rekomendasi untuk penerapan SMOTE dalam aplikasi klasifikasi lainnya. Diagram ini menunjukkan pendekatan sistematis dalam penelitian yang bertujuan untuk mengatasi tantangan ketidakseimbangan kelas menggunakan SMOTE dan algoritma KNN. Dengan pendekatan ini, diharapkan model klasifikasi yang dibangun dapat menghasilkan prediksi yang lebih akurat dan andal.

A. Pembagian Dataset dan *Preprocessing*

Dataset yang digunakan dalam penelitian ini terdiri dari total 619 data yang didistribusikan ke dalam tiga bagian utama: data pelatihan, validation, dan pengujian. Sebagian besar Dataset, yaitu 363 data atau sekitar 59%, dialokasikan untuk tahap pelatihan. Data ini digunakan untuk melatih model, sehingga model dapat mempelajari pola-pola yang ada dalam data dan membangun kemampuan untuk melakukan prediksi secara akurat. Tahap pelatihan ini memerlukan proporsi data yang besar untuk memastikan model dapat memahami representasi data dengan baik. Selain itu, sebanyak 156 data atau sekitar 25% digunakan sebagai data validation. Data validation berfungsi untuk mengevaluasi performa model selama pelatihan, memastikan model tidak mengalami masalah overfitting, serta membantu menentukan parameter optimal, seperti nilai k dalam algoritma KNN. Sementara itu, bagian terakhir dari Dataset, yaitu 100 data atau sekitar 16%, dialokasikan untuk pengujian. Data pengujian ini sepenuhnya dipisahkan dari proses pelatihan dan validation untuk mengevaluasi kemampuan generalisasi model terhadap data baru. Dengan proporsi pembagian seperti ini, proses pelatihan, validation, dan pengujian model dapat dilakukan secara sistematis untuk memastikan bahwa model yang dibangun tidak hanya bekerja dengan baik pada data pelatihan tetapi juga mampu memberikan prediksi yang akurat pada data yang belum pernah dilihat sebelumnya. Pendekatan ini merupakan praktik terbaik dalam pembelajaran mesin untuk menghasilkan model yang andal dan memiliki performa tinggi.

Dataset yang digunakan dalam penelitian ini terdiri dari dua kelas komentar, yaitu kelas positif dan negatif. Kedua kelas ini mewakili sentimen dari komentar yang dianalisis. Komentar positif mencerminkan opini atau pernyataan yang bersifat mendukung atau menyenangkan, sedangkan komentar negatif mencerminkan opini atau pernyataan yang bersifat kritis atau kurang mendukung. Distribusi kelas dalam Dataset dapat dilihat pada Tabel 1.

TABEL 1
DISTIBUSI DATA SET

Data	Positif	Negatif	Total	Total Dataset
Training	215	148	363	619
Validation	92	64	156	
Testing	50	50	100	

Berdasarkan tabel 1 distribusi data kelas positif dan negatif tidak merata. Distribusi data yang tidak merata seperti ini memengaruhi kinerja model, terutama dalam mengenali pola pada kelas dengan jumlah data yang lebih kecil (kelas minoritas). Oleh karena itu, penelitian ini menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) untuk menyeimbangkan jumlah data antar kelas, sehingga model dapat mempelajari kedua kelas secara proporsional dan memberikan hasil prediksi yang lebih akurat.

B. Training dan Validaso Model KNN

Pada tahap ini, model KNN (*K-Nearest Neighbors*) dibangun melalui proses pelatihan (*training*) dan *validation* menggunakan nilai kkk dari 1 hingga 25. Nilai k ini merepresentasikan jumlah tetangga terdekat yang digunakan oleh model untuk menentukan kelas dari data uji. Proses ini bertujuan untuk mengevaluasi performa model dalam memprediksi data dengan tingkat akurasi yang tinggi. Pelatihan dilakukan dengan menggunakan data yang telah melalui pra-pemrosesan sebelumnya, termasuk penerapan SMOTE untuk menyeimbangkan distribusi kelas. Setiap iterasi model dengan berbagai nilai kkk diuji untuk menemukan parameter terbaik yang menghasilkan performa optimal.

Hasil dari proses pelatihan dan *validation* kemudian dievaluasi menggunakan matriks evaluasi, yaitu *Confusion Matrix*. Matriks ini memberikan gambaran detail tentang jumlah prediksi benar dan salah untuk setiap kelas, baik positif maupun negatif. Dari *Confusion Matrix*, dihitung empat metrik evaluasi utama, yaitu Akurasi, Presisi, *Recall*, dan *F1-Score*.

C. Testing menggunakan K terbaik

Setelah melakukan proses training dan *validation* pada model KNN dengan nilai k dari 1 hingga 25, model dengan nilai k terbaik yang memberikan performa paling optimal pada tahap *validation* dipilih untuk digunakan pada tahap berikutnya. Model dengan nilai k terbaik ini dianggap mampu mengenali pola data dengan akurasi tinggi berdasarkan hasil evaluasi pada data latih dan *validation*. Setelah itu, model yang telah terpilih diuji lebih lanjut menggunakan *Dataset* testing untuk mengevaluasi kemampuan generalisasinya terhadap data yang tidak terlihat sebelumnya. *Dataset* testing ini terdiri dari 50 komentar positif dan 50 komentar negatif, sehingga distribusi kelasnya seimbang.

Proses testing bertujuan untuk mengukur seberapa baik model dapat memprediksi kelas positif dan negatif dengan benar pada data baru, yang tidak digunakan selama proses pelatihan maupun *validation*.

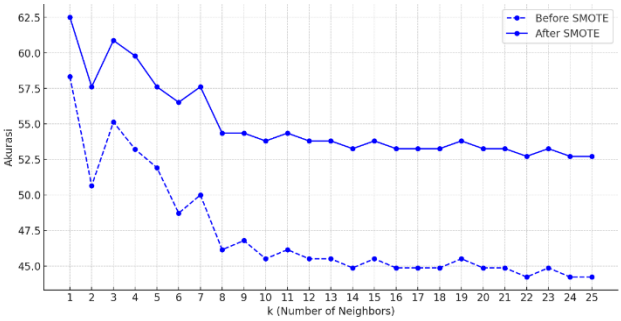
IV. HASIL DAN PEMBAHASAN

A. Hasil *Training* dan *Validation*

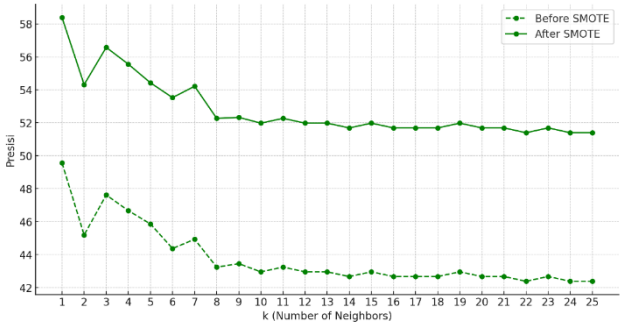
Pengujian model klasifiaksi dilakukan dengan menggunakan nilai K-1 sampai dengan K-25 dengan menampilkan hasil akurasi, presisi, *Recall* dan *F1 Score* dan dibandingkan dengan perbandingan data set orsinil sebelum menggunakan SMOTE dan *Dataset* seteleh menggunakan SMOTE. Hasil pengujian akurasi, presisi, *recall* dan *F1 Score* sebelum data set menggunakan SMOTE dan setelah data set menggunakan SMOTE dapat dilihat pada Tabel 2 dan Gambar 3 sampai Gambar 6.

TABEL 2
HASIL AKURASI, PRESISI , RECALL DAN F1 SCORE PADA PENGUJIAN K-1 SAMPAI K-25

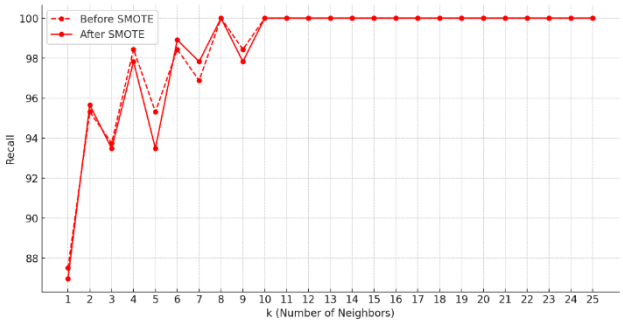
K-n	Sebelum SMOTE				Setelah SMOTE			
	Akurasi	**Presisi**	**Recall**	***F1-Score***	**Akurasi**	**Presisi**	**Recall**	***F1-Score***
1	58,33%	49,56%	87,50%	63,28 %	62,50%	58,39%	86,96%	69,87 %
2	50,64%	45,19%	95,31%	61,31 %	57,61%	54,32%	95,65%	69,29 %
3	55,13%	47,62%	93,75%	63,16 %	60,87%	56,58%	93,48%	70,49 %
4	53,21%	46,67%	98,44%	63,32 %	59,78%	55,56%	97,83%	70,87 %
5	51,92%	45,86%	95,31%	61,93 %	57,61%	54,43%	93,48%	68,80 %
6	48,72%	44,37%	98,44%	61,17 %	56,52%	53,53%	98,91%	69,47 %
7	50,00%	44,93%	96,88%	61,39 %	57,61%	54,22%	97,83%	69,77 %
8	46,15%	43,24%	100,00%	60,38 %	54,35%	52,27%	100,00%	68,66 %
9	46,79%	43,45%	98,44%	60,29 %	54,35%	52,33%	97,83%	68,18 %
10	45,51%	42,95%	100,00%	60,09 %	53,80%	51,98%	100,00%	68,40 %
11	46,15%	43,24%	100,00%	60,38 %	54,35%	52,27%	100,00%	68,66 %
12	45,51%	42,95%	100,00%	60,09 %	53,80%	51,98%	100,00%	68,40 %
13	45,51%	42,95%	100,00%	60,09 %	53,80%	51,98%	100,00%	68,40 %
14	44,87%	42,67%	100,00%	59,81 %	53,26%	51,69%	100,00%	68,15 %
15	45,51%	42,95%	100,00%	60,09 %	53,80%	51,98%	100,00%	68,40 %
16	44,87%	42,67%	100,00%	59,81 %	53,26%	51,69%	100,00%	68,15 %
17	44,87%	42,67%	100,00%	59,81 %	53,26%	51,69%	100,00%	68,15 %
18	44,87%	42,67%	100,00%	59,81 %	53,26%	51,69%	100,00%	68,15 %
19	45,51%	42,95%	100,00%	60,09 %	53,80%	51,98%	100,00%	68,40 %
20	44,87%	42,67%	100,00%	59,81 %	53,26%	51,69%	100,00%	68,15 %
21	44,87%	42,67%	100,00%	59,81 %	53,26%	51,69%	100,00%	68,15 %
22	44,23%	42,38%	100,00%	59,53 %	52,72%	51,40%	100,00%	67,90 %
23	44,87%	42,67%	100,00%	59,81 %	53,26%	51,69%	100,00%	68,15 %
24	44,23%	42,38%	100,00%	59,53 %	52,72%	51,40%	100,00%	67,90 %
25	44,23%	42,38%	100,00%	59,53 %	52,72%	51,40%	100,00%	67,90 %



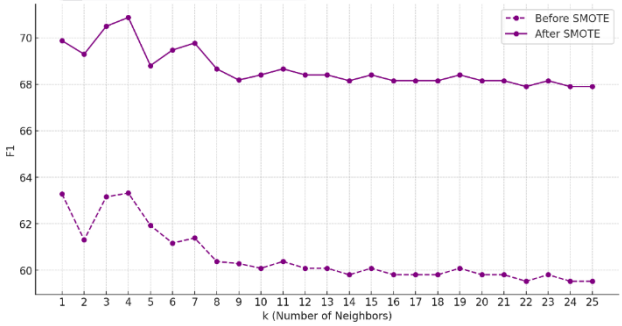
GAMBAR 3
Hasil Akurasi Sebelum dan Sesudah SMOTE



GAMBAR 4
Hasil Presisi Sebelum dan Sesudah SMOTE



GAMBAR 5
Hasil Recall Sebelum dan Sesudah SMOTE



GAMBAR 6
Hasil F1-Scorel Sebelum dan Sesudah SMOTE

Berdasarkan hasil pengujian, nilai k yang berbeda memberikan variasi pada akurasi, presisi, recall *dan* F1 Score dari model KNN Grafik akurasi menunjukkan perbandingan kinerja model KNN sebelum dan setelah SMOTE diterapkan pada berbagai nilai k dari 1 hingga 25. Sebelum SMOTE, akurasi model cenderung lebih rendah dan mengalami fluktuasi yang signifikan, terutama pada nilai k kecil. Pada k = 1, akurasi dimulai dengan 58.33% dan menurun drastis

pada nilai k = 2 hingga sekitar 50.64%, kemudian terus menurun hingga mencapai titik terendah sekitar 44.23% pada k = 24 dan k = 25. Fluktuasi ini menunjukkan bahwa tanpa SMOTE, model menghadapi kesulitan dalam menangani ketidakseimbangan kelas, sehingga hasil prediksi kurang stabil.

Setelah SMOTE diterapkan, akurasi model meningkat secara signifikan dan menjadi lebih stabil. Pada k = 1, akurasi meningkat menjadi 62.50% dan tetap lebih tinggi dibandingkan sebelum SMOTE pada hampir semua nilai k. Akurasi setelah SMOTE juga cenderung lebih konsisten dengan fluktuasi yang lebih kecil dibandingkan sebelumnya. Peningkatan ini mengindikasikan bahwa SMOTE membantu model KNN mempelajari pola dari kedua kelas dengan lebih baik, sehingga menghasilkan prediksi yang lebih akurat. Stabilitas yang lebih baik menunjukkan bahwa distribusi data yang seimbang memungkinkan model menghasilkan performa yang lebih andal.

Pada Grafik presisi, sebelum SMOTE, presisi model dimulai dari 49.56% pada k = 1, kemudian menurun secara bertahap dengan fluktuasi yang besar hingga mencapai titik terendah di sekitar 42.38% pada k = 24 dan k = 25. Penurunan ini menunjukkan bahwa tanpa SMOTE, model sering salah mengklasifikasikan data negatif sebagai positif, karena bias terhadap kelas mayoritas (positif).

Setelah SMOTE diterapkan, presisi model meningkat secara signifikan, dimulai dari 58.39% pada k = 1 dan tetap lebih tinggi dibandingkan presisi sebelum SMOTE pada semua nilai k. Peningkatan ini mencerminkan bahwa model menjadi lebih andal dalam mengklasifikasikan data positif dengan benar, mengurangi kesalahan prediksi positif yang salah. Selain itu, grafik presisi setelah SMOTE menunjukkan stabilitas yang lebih baik dengan fluktuasi yang lebih kecil, terutama pada k > 10, yang menunjukkan bahwa SMOTE membantu memperbaiki representasi kelas minoritas dalam data latih, sehingga model dapat memahami pola dengan lebih baik.

Sedangkan pada grafik *recall* menunjukkan kemampuan model dalam mendeteksi semua data positif di *Dataset*. Sebelum SMOTE, recall dimulai dengan nilai tinggi sebesar 87.50% pada k = 1 dan meningkat tajam hingga 100.00% pada k = 8 ke atas. Recall yang tinggi ini mencerminkan bahwa model mampu mendeteksi hampir semua data positif, meskipun sering kali disertai dengan prediksi positif yang salah, seperti yang terlihat dari presisi yang rendah sebelum SMOTE.

Setelah SMOTE, recall tetap tinggi dengan pola yang stabil. Nilai recall dimulai dari 86.96% pada k = 1 dan mencapai 100.00% pada k = 8 ke atas, sama seperti sebelum SMOTE. Stabilitas recall setelah SMOTE menunjukkan bahwa model tetap mampu mendeteksi hampir semua data positif meskipun data telah diimbangi. Dengan kata lain, SMOTE berhasil menjaga kemampuan model dalam mendeteksi kelas positif tanpa mengorbankan kemampuan ini untuk meningkatkan presisi. Kombinasi recall yang tinggi dan presisi yang meningkat membuat model menjadi lebih seimbang dan andal.

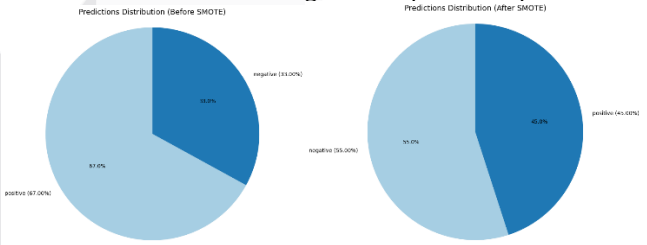
Grafik *F1-Score* menggabungkan informasi dari presisi dan recall, memberikan pandangan menyeluruh tentang keseimbangan performa model. Sebelum SMOTE, *F1-Score* dimulai dari 63.28% pada k = 1 dan mengalami penurunan bertahap dengan fluktuasi yang besar hingga mencapai titik terendah sekitar 59.53% pada k = 24 dan k = 25. Penurunan ini mengindikasikan bahwa ketidakseimbangan kelas menyebabkan model kurang mampu mempertahankan keseimbangan antara presisi dan recall.

Setelah SMOTE, *F1-Score* meningkat signifikan, dimulai dari 69.87% pada k = 1 dan tetap lebih tinggi dibandingkan sebelum SMOTE pada semua nilai k. Selain itu, *F1-Score* setelah SMOTE menunjukkan pola yang lebih stabil dengan fluktuasi yang lebih kecil, terutama pada nilai k > 10. Peningkatan ini menunjukkan bahwa SMOTE berhasil meningkatkan keseimbangan performa model, memastikan bahwa model tidak hanya mampu mendeteksi kelas positif tetapi juga menghasilkan prediksi positif yang lebih akurat. Berdasarkan nilai tersebut maka pada percobaan ini nilai K terbaik yang didapat ialah nilai K-1.

B. Hasil Testing mengunakan K Terbaik

Setelah Setelah melakukan *Training* dan *validation* pada model KNN dengan k1 sampai k25, dan didapatkan model dengan hasil terbaik. Selanjutnya model dengan hasil terbaik dilakukan testing dengan data Test 100 komentar dimana distribusi nya 50 komentar negative dan 50 komentar positif. Hasil perbandingan pada model dengan K terbaik (sebelum dan sesudah menggunakan SMOTE) dapat dilihat pada Gambar 7.

Berdasarkan Gambar 7, hasil prediksi yang ditampilkan pada diagram menggambarkan perbedaan distribusi prediksi model sebelum dan setelah penerapan SMOTE pada data pengujian. *Dataset* pengujian sebenarnya memiliki distribusi kelas yang seimbang, yaitu 50% komentar positif dan 50% komentar negatif, namun hasil prediksi model menunjukkan adanya ketidakseimbangan sebelum SMOTE diterapkan. Sebelum SMOTE, model cenderung mendominasi kelas positif dengan prediksi sebanyak 67% komentar positif dan hanya 33% komentar negatif. Ketimpangan ini mengindikasikan bahwa model lebih condong memprioritaskan kelas positif, kemungkinan besar akibat ketidakseimbangan kelas pada data pelatihan.



GAMBAR 7
Hasil *Tsting* dengan K Terbaik

Setelah SMOTE diterapkan, distribusi prediksi menjadi lebih mendekati proporsi yang seimbang, yaitu 45% komentar positif dan 55% komentar negatif. Meskipun hasil ini belum sepenuhnya mencerminkan distribusi nyata dari data pengujian, terlihat adanya perbaikan signifikan dalam kemampuan model untuk mengenali kelas negatif. Hal ini menunjukkan bahwa SMOTE berhasil membantu model mempelajari pola pada kelas minoritas di data pelatihan, sehingga prediksi pada kelas negatif meningkat secara signifikan.

V. KESIMPULAN

Berdasarkan pembahasan yang telah dilakukan, diperoleh beberapa kesimpulan utama. Pertama, SMOTE dapat diimplementasikan pada dataset yang tidak seimbang untuk meningkatkan akurasi model KNN. Kedua, nilai k=1 memberikan hasil terbaik dengan akurasi sebesar 62,50%, presisi 58,39%, recall 86,96%, dan F1-Score 69,87%, yang menunjukkan bahwa model KNN dengan k=1 setelah penerapan SMOTE memberikan performa paling optimal dalam menangani ketidakseimbangan kelas. Ketiga, analisis data menunjukkan bahwa penerapan SMOTE meningkatkan performa model KNN, dengan akurasi pada k=1 meningkat dari 58,33% menjadi 62,50%, presisi dari 49,56% menjadi 58,39%, dan F1-Score dari 63,28% menjadi 69,87%, sambil mempertahankan recall yang tinggi dari 87,50% menjadi 86,96%. Hal ini membuktikan bahwa SMOTE efektif dalam mengatasi ketidakseimbangan kelas dan meningkatkan kualitas prediksi model.

REFERENSI

[1] S. Sadya, "APJII: Pengguna Internet Indonesia 215,63 Juta pada 2022-2023," 2023. https://dataindonesia.id/internet/detail/apjii-pengguna-internet-indonesia-21563-juta-pada-20222023 (accessed Dec. 10, 2024).

[2] Irham Irham, Tasrif Tasrif, and Junaidin Junaidin, "Analisis Pemberitaan Media Online Kahaba.Net Dengan Bimakini.Com Tentang Masjid Terapung (Sebuah Framing Edisi Oktober 2017 – Januari 2018)," *J. Ilm. Tek. Inform. dan Komun.*, vol. 2, no. 1, pp. 88–98, 2022, doi: 10.55606/juitik.v2i1.267.

[3] A. N. Ulfah and M. K. Anam, "Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM)," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 1, pp. 1–10, 2020, doi: 10.35957/jatisi.v7i1.196.

[4] E. Sutoyo and M. A. Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 3, p. 379, 2020, doi: 10.26418/jp.v6i3.42896.

[5] N. Sulistiyowati and M. Jajuli, "Integrasi Naive Bayes Dengan Teknik Sampling Smote Untuk Menangani Data Tidak Seimbang," *Nuansa Inform.*, vol. 14, no. 1, p. 34, 2020, doi: 10.25134/nuansa.v14i1.2411.

[6] C. Cahyaningtyas, Y. Nataliani, and I. R. Widiasari, "Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE," *Aiti*, vol. 18, no. 2, pp. 173–184, 2021, doi: 10.24246/aiti.v18i2.173-184.

[7] G. A. Mursianto, I. M. Falih, M. Irfan, T. Sakinah, and D. S. Prasvita, "Perbandingan Metode Klasifikasi Random Forest dan XGBoost Serta Implementasi Teknik SMOTE pada Kasus Prediksi Hujan," *J. Senamika*, vol. 2, no. 2, pp. 41–50, 2021.

[8] M. Dewi, T. H. Saragih, and R. Herteno, "Penerapan SMOTE-NCL untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Jantung Koroner," *J. Inform. Polinema*, vol. 10, no. 1, pp. 27–34, 2023, doi: 10.33795/jip.v10i1.1394.

[9] A. Sinha, M. N. B. J. Naskar, M. Pandey, and S. S. Rautaray, "Text Classification Using ML Techniques: Comparative Analysis," *Proc. - 2022 OITS Int. Conf. Inf. Technol. OCIT 2022*, no. August, pp. 102–107, 2022, doi: 10.1109/OCIT56763.2022.00029.

[10] Z. Wan, "Text Classification: A Perspective of Deep Learning Methods," 2023, [Online]. Available: http://arxiv.org/abs/2309.13761

[11] Muhammad Zulqarnain *et al.*, "Text Classification Using Deep Learning Models: A Comparative Review," *Cloud Comput. Data Sci.*, no. March, pp. 80–96, 2023, doi: 10.37256/ccds.5120243528.

[12] N. Istiana and A. Mustafiril, "Perbandingan Metode Klasifikasi pada Data dengan Imbalance Class dan Missing Value," *J. Inform.*, vol. 10, no. 2, pp. 101–108, 2023, doi: 10.31294/inf.v10i2.15540.

[13] A. H. Wijaya, "Artificial Neural Network Untuk Memprediksi Beban Listrik Dengan Menggunakan Metode Backpropagation," *J. CoreIT*, vol. 5, no. 2, pp. 61–70, 2019.

[14] E. Grossi and M. Buscema, "Introduction to artificial neural networks," *Eur. J. Gastroenterol. Hepatol.*, vol. 19, no. 12, pp. 1046–1054, 2007, doi: 10.1097/MEG.0b013e3282f198a0.

[15] M. Faisal Riftiarrasyid, D. Arif Setyawan, and H. Maulana, "Klasifikasi Kesegaran Daging Sapi Menggunakan Metode Gray Level Cooccurrence Matrix dan DNN," *J. Ilm. Teknol. Inf. dan Robot.*, vol. 3, no. 2, pp. 34–38, 2021, doi: 10.33005/jifti.v3i2.65.

[16] L. Wiranda and M. Sadikin, "Penerapan Long Short Term Memory Pada Data Time Series Untuk Memprediksi Penjualan Produk Pt. Metiska Farma," *J. Nas. Pendidik. Tek. Inform.*, vol. 8, no. 3, pp. 184–196, 2019.

[17] E. Golafshani, A. A. Chiniforush, P. Zandifaez, and T. Ngo, "An artificial intelligence framework for predicting operational energy consumption in office buildings," *Energy Build.*, vol. 317, no. January, p. 114409, 2024, doi: 10.1016/j.enbuild.2024.114409.

[18] A. Pulver and S. Lyu, "LSTM with working memory," in *Proceedings of the International Joint Conference on Neural Networks*, 2017, pp. 845–851. doi: 10.1109/IJCNN.2017.7965940.

[19] R. G. Khalkar, A. S. Dikhit, and A. Goel, "Handwritten Text Recognition using Deep Learning (CNN & RNN)," *Iarjset*, vol. 8, no. 6, pp. 870–881, 2021, doi: 10.17148/iarjset.2021.86148.

[20] D. Koblah *et al.*, "A Survey and Perspective on Artificial Intelligence for Security-Aware Electronic Design Automation," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 28, no. 2, pp. 1–55, 2023, doi: 10.1145/3563391.

[21] P. R. Sihombing and A. M. Arsani, "Comparison of ML Methods in Classifying Poverty in Indonesia in 2018 Perbandingan Metode ML Dalam Klasifikasi Kemiskinan Di Indonesia Tahun 2018," *J. Tek. Inform.*, vol. 2, no. 1, pp. 51–56, 2021.

[22] S. Nurmani, A. Darmawahyuni, A. I. Sapitri, M. N. Rachmatullah, Firdaus, and B. Tutuko, "Peengenalan Deep Learning dan Implementasinya," p. 137, 2021.

[23] A. A. Pratama, Y. Yohanie, F. Panduman, D. K. Basuki, and S. Sukaridhoto, "Edge Computing Implementation for Action Recognition Systems," *Sci. J. Informatics*, vol. 7, no. 2, pp. 2407–7658, 2020, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/sji

[24] R. Rahmadini, Enjel Erika LorencisLubis, Aji Priansyah, Yolanda R.W.N, and Tuti Meutia, "Penerapan Data Mining Untuk Memprediksi Harga Bahan Pangan Di Indonesia Menggunakan Algoritma K-Nearest Neighbor," *J. Mhs. Akunt. Samudra*, vol. 4, no. 4, pp. 223–235, 2023, doi: 10.33059/jmas.v4i4.7074.

[25] V. A. Nugroho, D. P. Adi, A. T. Wibowo, M. T. Sulistyono, and A. B. Gumelar, "Klasifikasi Jenis Pemeliharaan dan Perawatan Container Crane menggunakan Algoritma ML," *Matics*, vol. 13, no. 1, pp. 21–27, 2021, doi: 10.18860/mat.v13i1.11525.

[26] A. B. Nugraha and A. Romadhony, "Identification of 10 Regional Indonesian Languages Using ML," *Sinkron*, vol. 8, no. 4, pp. 2203–2214, 2023, doi: 10.33395/sinkron.v8i4.12989.

[27] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada *Dataset* Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.