ISSN : 2527-5836
e-ISSN : 2528-0074

Vol. 9 No. 2, Mei 2024

JISKa

Jurnal Informatika Sunan Kalijaga

Jurusan Teknik Informatika
Fakultas Sains dan Teknologi
UIN Sunan Kalijaga Yogyakarta

Tim Pengelola JISKa (Jurnal Informatika Sunan Kalijaga)
Edisi Mei 2024

Ketua Editor (*Editor in Chief*)

Muhammad Taufiq Nuruzzaman, Ph.D. (UIN Sunan Kalijaga Yogyakarta, Indonesia)

Dewan Editor (*Editorial Board*)

1. Dr. Aang Subiyakto (UIN Syarif Hidayatullah Jakarta, Indonesia)
2. Andang Sunarto, Ph.D. (IAIN Bengkulu, Indonesia)
3. Dr. Hamdani (Universitas Mulawarman Samarinda, Indonesia)
4. Nashrul Hakiem, Ph.D. (UIN Syarif Hidayatullah Jakarta, Indonesia)
5. Noor Akhmad Setiawan, Ph.D. (Universitas Gadjah Mada, Indonesia)

Editor Bahasa dan Layout (*Copy Editor and Layout Editor*)

Sekar Minati, S.Kom. (UIN Sunan Kalijaga Yogyakarta, Indonesia)

Tim Teknologi Informasi (*Journal Manager and Technical Support*)

1. Eko Hadi Gunawan, M.Eng. (UIN Sunan Kalijaga Yogyakarta, Indonesia)
2. Muhammad Galih Wonoseto, M.T. (UIN Sunan Kalijaga Yogyakarta, Indonesia)

Mitra Bestari (*Reviewer*)

Reviewer Internasional:

1. Ardiansyah Musa Efendi, Ph.D. (Singapore Chipset Algorithm Design Lab, Huawei, Singapore)
2. Dr.Eng. M. Muhammad Syafrudin (Sejong University, Korea Selatan)
3. Dr.Eng. M. Alex Syaekhoni (Who's Good, Korea Selatan)
4. Norma Latif Fitriyani, M.Sc. (Sejong University, Korea Selatan)

Reviewer Nasional:

1. Dr. Ir. Agung Fatwanto (UIN Sunan Kalijaga Yogyakarta, Indonesia)
2. Agus Mulyanto, S.Si., M.Kom., ASEAN Eng. (UIN Sunan Kalijaga Yogyakarta, Indonesia)
3. Ahmad Fathan Hidayatullah, M.Cs. (Universitas Islam Indonesia Yogyakarta, Indonesia)
4. Alam Rahmatulloh, M.T. (Universitas Siliwangi Tasikmalaya, Indonesia)
5. Anggi Rizky Windra Putri, M.Kom. (Universitas Aisyiyah Yogyakarta, Indonesia)
6. Dr. Ir. Bambang Sugiantoro (UIN Sunan Kalijaga Yogyakarta, Indonesia)
7. Dr. Enny Itje Sela (Universitas Teknologi Yogyakarta, Indonesia)
8. Dr.Eng. Ganjar Alfian (Universitas Gadjah Mada, Indonesia)
9. Mandahadi Kusuma, M.Eng. (UIN Sunan Kalijaga Yogyakarta, Indonesia)
10. Maria Ulfa Siregar, Ph.D. (UIN Sunan Kalijaga Yogyakarta, Indonesia)
11. Muhammad Dzulfikar Fauzi, M.Cs. (Telkom University Surabaya, Indonesia)
12. Muhammad Habibi, M.Cs. (Universitas Jenderal Achmad Yani Yogyakarta, Indonesia)
13. Muhammad Rifqi Maarif, M.Eng. (Universitas Jenderal Achmad Yani Yogyakarta, Indonesia)
14. Niki Min Hidayati Robbi, M.Eng. (Universitas Gadjah Mada, Indonesia)
15. Prof. Dr. Hj. Okfalisa, S.T., M.Sc. (UIN Sultan Syarif Kasim Riau, Indonesia)
16. Oman Somantri, M.Kom. (Politeknik Negeri Cilacap, Indonesia)
17. Puguh Jayadi, M.Kom. (Universitas PGRI Madiun, Indonesia)
18. Puji Winar Cahyo, M.Cs. (Universitas Jenderal Achmad Yani Yogyakarta, Indonesia)
19. Qorry Aina Fitroh, M.Kom. (UIN Sunan Kalijaga Yogyakarta, Indonesia)
20. Ridho Surya Kusuma, M.Kom. (Universitas Siber Muhammadiyah, Yogyakarta, Indonesia)
21. Dr. Shofwatul Uyun (UIN Sunan Kalijaga Yogyakarta, Indonesia)
22. Dr. Ir. Sumarsono, M.Kom. (UIN Sunan Kalijaga Yogyakarta, Indonesia)
23. Dr.Eng. Sunu Wibirama, M.Eng. (Universitas Gadjah Mada, Indonesia)
24. Tundo, M.Kom. (Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika (STIKOM CKI), Indonesia)
25. Yudistira Dwi Wardhana Asnar, Ph.D. (Institut Teknologi Bandung, Indonesia)

ISSN : 2527-5836

e-ISSN: 2528-0074

JISKa (Jurnal Informatika Sunan Kalijaga)

Vol. 9, No. 2, MEI 2024

DAFTAR ISI

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024, Pp. 79 – 93
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Analisa Jejaring Sosial Terhadap Fenomena *Cyberbullying* Fandom K-Pop pada Sosial Media Twitter

Mohammad Iqbal Ghufron [(1)*], Endang Supriyati [(2)], Tri Listyorini [(3)]
Teknik Informatika, Fakultas Teknik, Universitas Muria Kudus, Kudus
e-mail : 201951009@std.umk.ac.id, {endang.supriyati,trilistyorini}@umk.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 27 Agustus 2023, direvisi 23 Februari 2024, diterima 26 Februari 2024, dan dipublikasikan 25 Mei 2024.

Abstract

This study examines cyberbullying among K-pop fandoms through social network analysis (SNA) using data from Twitter, a social media platform. The phenomenon of K-pop gaining global popularity also brings negative impacts, such as cyberbullying, which can affect the psychological well-being of victims. Using R Studio and Gephi analysis tools, this study applied centrality values, including degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality, to identify influential accounts in the spread of the cyberbullying phenomenon. This analysis provides insight into the interaction and influence between Twitter user accounts in the context of cyberbullying. The main objective of this research is to paint a picture of the cyberbullying phenomenon involving various K-pop fandoms and identify the accounts that play an essential role in the related communication network.

Keywords: K-pop, Cyberbullying, Centrality, Fandom, SNA

Abstrak

Penelitian ini mengkaji fenomena *cyberbullying* di kalangan *fandom* K-pop melalui analisis jejaring sosial (SNA) menggunakan data dari media sosial Twitter. Fenomena K-pop yang meraih popularitas global turut membawa dampak negatif, seperti *cyberbullying*, yang dapat berdampak pada kesejahteraan psikologis korban. Tujuan utama penelitian ini adalah menggambarkan gambaran fenomena *cyberbullying* yang melibatkan berbagai *fandom* K-pop serta mengidentifikasi akun-akun yang memiliki peran penting dalam jaringan komunikasi terkait. Dengan menggunakan alat analisis R Studio dan *Gephi*, penelitian ini menerapkan nilai *centrality*, termasuk *degree centrality, betweenness centrality, closeness centrality,* dan *eigenvector centrality*, untuk mengidentifikasi akun-akun berpengaruh dalam penyebaran fenomena *cyberbullying*. Dampak dari penelitian ini yang paling berpengaruh dalam fenomena *cyberbullying* adalah pada akun "*tanyakanrl*". Pada akun tersebut paling banyak dilakukan retweet. Hasil analisis ini memberikan wawasan tentang interaksi dan pengaruh antara akun-akun pengguna Twitter dalam konteks fenomena *cyberbullying* tersebut.

Kata Kunci: K-pop, *Cyberbullying*, *Centrality*, *Fandom*, SNA

1. PENDAHULUAN

Musik K-pop telah menjadi fenomena global yang mendunia, menarik perhatian jutaan penggemar dari berbagai belahan dunia. K-pop, atau Korean *Pop*, merupakan aliran musik populer yang berasal dari Korea Selatan dan mencakup berbagai gaya dan *genre*, mulai dari *pop*, *hip-hop*, R&B, hingga *rock* dan *dance*. K-pop tidak hanya menawarkan pengalaman musik yang unik, tetapi juga memadukan elemen budaya dan estetika Korea, menciptakan identitas visual dan artistik yang menarik dan memukau.

Sebagai bagian dari *Hallyu* atau "Gelombang Korea," K-pop telah membawa pengaruh budaya Korea Selatan ke seluruh dunia, termasuk musik, drama televisi, film, *fashion*, dan makanan. *Idol* grup seperti TWICE, BTS, BLACKPINK, dan EXO, serta solo *artist* seperti Taeyeon, IU, Jessy, Sunmi, dan D.O, telah meraih popularitas yang luar biasa di Asia, termasuk Indonesia, dan bahkan menyebar ke Amerika Serikat dan Eropa. Kecanduan akan drama cinta Asia, terutama

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

drama Korea, juga turut memperkenalkan banyak anak muda pada dunia K-Pop, menambah popularitas fenomena ini di kalangan penggemar.

Di era media sosial, *fandom* K-pop semakin marak dan menjadi kekuatan besar dalam mendukung dan mempromosikan idolanya. *Fandom* mengacu pada sekelompok penggemar yang mengidolakan grup atau idola K-pop tertentu dan berbagi informasi serta dukungan melalui berbagai platform media sosial, terutama Twitter (Tionardus & Setuningsih, 2022). Media sosial ini memberikan kesempatan bagi penggemar untuk berperan sebagai produsen konten, membagikan informasi, berkomunikasi, dan memberikan dukungan kepada idola mereka.

Namun, di balik fenomena K-pop yang menyenangkan ini, ada sisi gelap yang perlu diperhatikan, yaitu fenomena *cyberbullying*. *Cyberbullying* adalah tindakan *bully* yang dilakukan melalui media sosial, termasuk di antara fandom K-pop (Gradinger et al., 2010). Pada dasarnya *bullying* melibatkan dua faktor yaitu pelaku *bully* atau pelaku intimidasi dan korbannya, pelaku intimidasi menindas korban secara fisik, verbal atau lainnya untuk mendapatkan rasa superioritas dan kekuasaan (Donegan, 2012). Fenomena ini dapat menyebabkan efek negatif pada psikologi korbannya, termasuk antara lain depresi atau stres, karena seringnya penghinaan, frekuensi perasaan sedih dan melankolis yang menyebabkan stres dan depresi, meningkatkan jumlah korban serta efeknya berjangka panjang (Irawan, 2018). Oleh karena itu, penting untuk memahami lebih dalam tentang fenomena *cyberbullying* yang melibatkan fandom K-pop dan identifikasi akun-akun yang berpengaruh dalam menyebarkan informasi negatif atau menyebarkan tindakan *cyberbullying*.

Penelitian ini menggunakan metode *Social Network Analysis* (SNA) untuk menganalisis data dari media sosial Twitter dan mencari nilai *centrality* dari akun-akun yang terlibat dalam fenomena *cyberbullying* di kalangan *fandom* K-pop. Metode SNA diterapkan karena untuk mencari yang paling berpengaruh di dalam jejaring paling tepat menggunakan metode *Social Network Analysis* (SNA). Penelitian ini juga berfokus pada analisis dinamika *cyberbullying* di komunitas fandom K-pop di Twitter (Lee & Jang, 2020). Dengan demikian, diharapkan penelitian ini dapat memberikan gambaran yang lebih jelas tentang seberapa besar dampak *cyberbullying* di kalangan penggemar K-pop dan mengidentifikasi akun-akun yang perlu diperhatikan dalam menciptakan lingkungan *online* yang lebih aman dan positif.

2. METODE PENELITIAN

Penelitian ini menggunakan metode *text mining* dan *Social Network Analysis* (SNA). proses ekstraksi informasi yang berguna dan bermakna dari teks yang tidak terstruktur. *Text Mining* merupakan teknik yang melibatkan penggunaan algoritma dan metode komputasional untuk menganalisis, mengklasifikasikan, dan mengekstrak pola, pengetahuan, atau wawasan dari teks yang ada. Menurut Ronen Feldman dan James Sanger, *text mining* adalah suatu proses menggali informasi di mana orang menggunakan alat analisis untuk berinteraksi dengan sekumpulan dokumen (Akbar, 2021). Sedangkan SNA adalah metode untuk mempelajari dan menganalisis hubungan antara individu, kelompok, atau entitas dalam suatu jaringan sosial. Dengan menggunakan teori graf, analisis jaringan sosial (SNA) akan membantu memudahkan dalam menggambarkan atau memvisualisasikan pola struktur jaringan ikatan sosial dalam suatu kelompok untuk menentukan hubungan atau hubungan antar individu (Tsvetovat & Kouznetsov, 2011).

Penelitian ini menerapkan pendekatan *Survey Online* untuk menganalisis teks yang dihasilkan oleh pengguna Twitter dalam media sosial. Penelitian analisis teks media sosial ini fokus pada data kuantitatif yang mencakup kategori jumlah *Tweet*, *Retweet*, dan penggunaan tagar "*cyberbullying*, *kpop*, dan *fandom*" oleh pengguna Twitter. Data tersebut mencerminkan interaksi dan kontribusi pengguna Twitter dalam konteks penelitian ini. Penelitian ini juga menggunakan pendekatan *sentiment analysis* terhadap *tweet* yang terkait dengan fandom K-pop, khususnya fokus pada insiden *cyberbullying* (Kim & Park, 2021).

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Dalam penelitian ini, penulis menggunakan level analisis pada aktor untuk mengamati interaksi yang terjadi antara para pengguna. Fokusnya adalah melihat bagaimana pengguna Twitter *(actors/nodes)* berinteraksi dalam pembahasan mengenai topik *Cyberbullying fandom* K-pop. Seberapa kuat dan seperti apa hubungan terjadi, apakah hubungan satu arah atau dua arah, bagaimana hubungan difasilitasi, dan melalui media apa hubungan terjadi. Aplikasi lain, seperti siapa yang memiliki hubungan (*ties*) terbanyak, siapa yang terisolasi dalam *network*, berapa jarak (*gap*) dan rentang (*length*) antar *nodes*, di mana terjadi *bottleneck*, siapa yang menjadi pemain penting, dan sebagainya, semuanya bergantung pada aktor yang terhubung.

Berikut jenis *centrality* individu yang paling umum:
1) *Degree centrality* : yang didefinisikan sebagai jumlah koneksi yang dimiliki sebuah *node*.
2) *Closeness centrality* : yang didefinisikan sebagai jarak rata-rata antara sebuah *node* dan semua *node* lain dalam jaringan. Ukuran ini menunjukkan seberapa dekat *node* ini dengan *node* lain. Orang yang lebih dekat dengan orang lain memiliki lebih banyak hubungan.
3) *Betweenness centrality* : ukuran ini menunjukkan fungsi sebuah *node* sebagai *bottleneck*. Persimpangan semakin penting seiring dengan jumlah jalan yang harus melewatinya (jika tidak ada jalan alternatif).
4) *Eigenvector centrality* : Mengukur pentingnya sebuah simpul berdasarkan kualitas hubungannya dengan simpul-simpul lain yang juga penting.

2.1 Fokus Penelitian

Fokus penelitian merupakan titik pusat atau topik yang diteliti dalam suatu studi. Ini membantu menentukan ruang lingkup penelitian, pertanyaan penelitian, dan metode yang digunakan. Dengan fokus yang jelas, penulis dapat mendapatkan pemahaman yang lebih mendalam tentang topik tertentu, menemukan temuan baru, atau memecahkan masalah yang ada. Penelitian ini berfokus untuk menganalisis dan menentukan *betweenness centrality* untuk mengetahui aktor yang paling berpengaruh pada tweet yang berkaitan dengan *cyberbullying* K-Pop dan bagaimana interaksi antar *fandom* K-Pop dengan *tweetwar* dan bagaimana konten atau isi *twitwar* tersebut dapat digolongkan dalam tipe-tipe *cyberbullying* sesuai dengan judul penelitian yaitu "Analisa Jejaring Sosial Terhadap Fenomena *Cyberbullying Fandom* K-Pop Pada Sosial Media Twitter".

2.2 Unit Analisis

2.2.1 *Tweet* dan *Retweet*

Tweet adalah pesan atau status pendek yang dibagikan oleh pengguna di platform media sosial Twitter untuk memberitahukan kondisi atau untuk menginformasikan kepada *follower* tentang suatu perkembangan yang sedang dialami oleh seseorang. Pengguna Twitter dapat membuat *tweet* untuk berbagi pemikiran, informasi, gambar, video, tautan, atau tagar terkait topik tertentu. *Retweet* adalah tindakan ketika pengguna Twitter membagikan kembali atau menyebarkan *tweet* yang telah di*posting* oleh pengguna lain. Dengan *retweet*, pengguna Twitter dapat memperluas jangkauan dan visibilitas *tweet* tersebut kepada pengikut mereka sendiri. *Retweet* sering digunakan untuk menyebarkan informasi menarik, pemikiran, atau konten yang dianggap relevan oleh pengguna Twitter.

2.2.2 Komentar dan *Reply*

Reply atau balasan adalah tanggapan langsung yang diberikan oleh pengguna terhadap komentar atau pesan yang telah diterima. Ini memungkinkan pengguna untuk terlibat dalam percakapan yang lebih mendalam dan berkelanjutan dengan pengguna lain. Komentar dan *reply* adalah cara yang penting dalam membangun interaksi sosial dan memperluas diskusi di platform media sosial. Mereka memungkinkan pengguna untuk saling berkomunikasi, bertukar ide, memberikan dukungan, atau mengajukan pertanyaan dalam konteks yang lebih terstruktur dan terorganisir.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

2.2.3 *Hashtag* atau Tagar (#)

Tagar "#" (tanda pagar) sering digunakan di media sosial, terutama di platform seperti Twitter, untuk mengidentifikasi dan mengelompokkan pesan atau postingan yang berkaitan dengan topik tertentu. Tagar biasanya digunakan tanpa spasi di antara kata-kata atau frasa. Tagar memungkinkan pengguna media sosial untuk menemukan konten yang relevan atau terkait dengan topik yang mereka minati. Tagar juga membantu dalam mengorganisir dan memperluas jangkauan pesan atau informasi yang ingin disampaikan oleh pengguna. Selain menjadi alat pencarian dan pengelompokan, tagar juga sering digunakan dalam kampanye, acara, atau gerakan sosial untuk menyatukan orang-orang di sekitar topik atau tujuan tertentu. Tagar dapat menjadi cara yang efektif untuk menghubungkan orang-orang dengan minat yang sama dan memperluas jangkauan pesan atau pesan yang ingin disampaikan di media sosial. Sehingga dalam hal ini, tagar yang akan digunakan dalam penelitian yaitu #*cyberbullying*, #*kpop*, dan #*fandom*.

2.3 Jenis dan Sumber Data

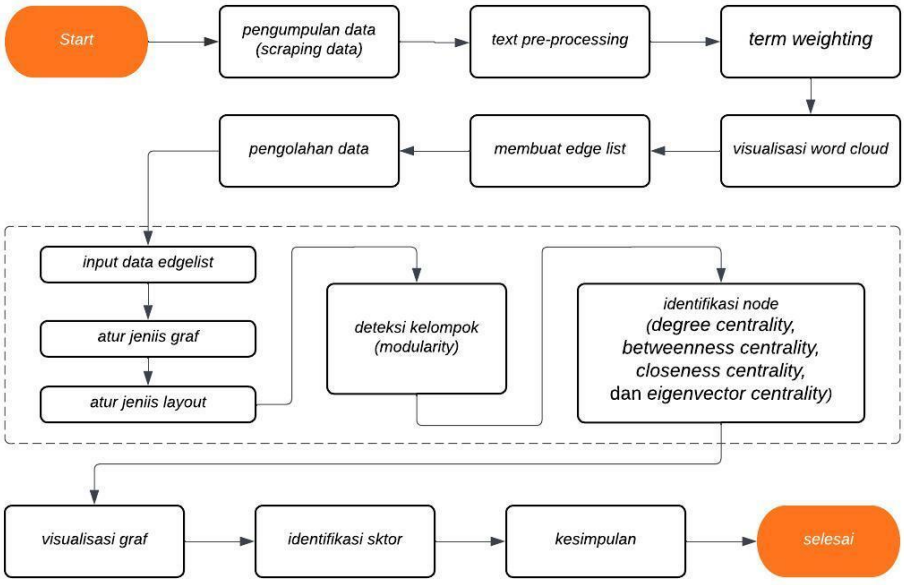
2.3.1 Data Primer

Data primer merujuk pada data yang dikumpulkan langsung dari sumber pertama dalam rangka penelitian atau tujuan tertentu. Dalam hal ini, data diperoleh dari sumber utama yaitu *tweet*, *retweet*, dan *hashtag* "*cyberbullying*, *kpop*, dan *fandom*" di Twitter. *Google collab* dengan menggunakan *tweet harvest* digunakan dalam proses *crawling* data dan mengumpulkan data pengguna Twitter aktif yang terlibat dalam topik *cyberbullying* K-Pop.

2.3.2 Data Sekunder

Data Sekunder merujuk pada data yang telah dikumpulkan oleh pihak lain atau dalam konteks yang berbeda sebelumnya. Pada tahap ini, peneliti mengumpulkan data dengan menelusuri artikel, jurnal, dan penelitian terkait untuk melengkapi kebutuhan penelitian. Data tersebut diperoleh dari sumber sekunder yang telah dikumpulkan sebelumnya oleh orang lain.

2.4 Tahap Penelitian



Gambar 1 Flowchart Tahap Penelitian

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Penelitian ini menggambarkan langkah-langkah yang diambil dalam pengumpulan dan analisa data terkait topik *cyberbully kpop* yang terlihat pada Gambar 1. Pertama, data dikumpulkan dari platform jejaring sosial media Twitter, menggunakan teknik *web scraping* dengan kata kunci "*cyberbullying*", "*kpop*", dan "*fandom*". *Tools* yang digunakan meliputi Google Colab dengan *library tweet harvest* dan Rstudio. Setelah terkumpul, data melalui proses *text preprocessing*, yaitu suatu prosedur untuk memilih data teks yang lebih terstruktur melalui berbagai langkah seperti *case folding, tokenizing, filtering*, dan *stemming* (Tineges & Davita, 2021). Selanjutnya, pembobotan kata atau *term weighting* diterapkan untuk memberikan nilai numerik pada kata-kata dalam teks, mencerminkan yang merepresentasikan pentingnya kata dalam teks tersebut. Visualisasi data dalam bentuk *word cloud* dilakukan dengan bantuan paket *wordcloud* dan *RcolorBrewer*.

Langkah berikutnya adalah membuat data *edge list* menggunakan *package ggplot2* dan *tidyverse* di R, khususnya untuk interaksi *reply* dan *retweet* di Twitter. Dua jenis *node* diperlukan: yang melakukan balasan (*reply*) dan yang menerima balasan. Analisis jaringan sosial (SNA) dilakukan dengan *software Gephi*. Prosesnya termasuk *input* data *edge list*, memilih jenis *graph undirected* sehingga mendapatkan jumlah *nodes* dan *edges*, memilih *layout* visualisasi graf, menghitung *modularity*, *degree centrality, betweenness centrality, eigenvector centrality*, dan *closeness centrality.*

Dalam teori graf dan *network analysis*, terdapat empat cara untuk mengukur *centrality*, yaitu dengan cara *menghitung degree centrality, betweeness centrality, closeness centrality, dan eigenvector centrality*. Pada penelitian ini akan digunakan dua cara perhitungan, yaitu *betweeness centrality* dan *closeness centrality*. *Betweenness centrality* adalah salah satu cara untuk mengukur *centrality* dalam suatu jaringan sosial. Berikut adalah rumus untuk menghitung nilai *betweenness centrality* setiap *node* dalam jaringan (Bader et al., 2007; Brandes, 2001).

Hasil analisis diekspresikan dalam bentuk visualisasi graf yang berbeda berdasarkan *modularity*. Identifikasi aktor yang paling berpengaruh dalam masing-masing ukuran *centrality* menjadi langkah penting berikutnya. Pada akhirnya, penelitian ini memberikan kesimpulan dan hasil dari penelitian *text mining* serta analisis SNA yang telah dilakukan, memberikan wawasan tentang pola interaksi dan aktor penting dalam topik yang diteliti.

3. HASIL DAN PEMBAHASAN

Dalam proses *scraping*, penulis menggunakan layanan *Google Collab* untuk menjalankan bahasa pemrograman *Python* dan menggunakan *library tweet harvest*. Hal ini dilakukan karena adanya kebijakan baru yang diterapkan oleh Elon Musk pada tanggal 2 Juli 2023, yang membatasi jumlah maksimal *tweet* yang dapat dibaca setiap hari untuk beberapa tipe pengguna. Proses *scraping* data Twitter dilakukan dengan menggunakan *keyword* "*kpop bully*" dengan bahasa Indonesia. Hasil dari proses *scraping* ini menghasilkan 739 *tweets* yang berkaitan dengan *keyword* tersebut. Dari proses *scraping* tersebut didapatkan hasil data yang dapat dilihat pada Tabel 1.

3.1 *Text Preprocessing*

Di sini data hasil dari proses *scraping* masih memiliki format yang tidak terstruktur, sehingga informasi di dalamnya sulit diekstrak secara langsung. Pada proses *scraping* menggunakan Tweet Harvest. Di mana Tweet Harvest merupakan *tools* yang digunakan untuk melakukan *crawling* data pada media sosial Twitter dengan menggunakan *Application Programming Interface* (API) (Yuniar et al., 2022). Oleh karenanya, diperlukan tahapan *text preprocessing* yang bertujuan untuk merapikan data agar memiliki format yang terstruktur. Tahap *text preprocessing* sangat penting dalam *text mining* karena membantu persiapan data agar dapat diolah lebih lanjut dengan lebih baik. Pada proses *preprocessing* terdiri dari beberapa tahapan, antara lain yaitu:

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Tabel 1 Contoh Data Hasil *Scraping*

No.	*Tweets*	*Username*	created_at
1	@jenjensante sampai ada twit sejahat itu loh dan semua twit bully lainnya mayoritas dari solo standnya kpop	onlyfreenbecky_	Tue Jul 18 15:49:39 +0000 2023
2	Kalah di bully Menang di kira curang, pake vpn lah, minta bantuan fandom lain lah (one person vs kpop cenah) di kira fandom kpop pro ke exol semua apa, mereka juga punya bias nya buat di vote elahh.. Big Love buat EXO + L aja❤ EXO VS EXOL sebenernya mah lu mah gak di ajak	NagaAsli_	Mon Jul 17 16:05:15 +0000 2023
3	Kpop itu ibarat lu boleh bully gua, boleh ngejatuhin gue, tp lu jangan pernah mengusik obat kebahagiaan gue	aletheafiani	Mon Jul 17 15:24:33 +0000 2023
4	tp emang baek kooyoung di bully abis abisan, gak cuma sama anak bp dan zb1 tapi juga sama idol idol kpop lain. salah sendiri sih..... jahat banget pas ngomong😭	01W00NGZ	Sun Jul 16 11:23:32 +0000 2023
5	@RuNext_base ga punya adab emg kebiasaan tukang bully disekolah, dibawa ke kpop	Cattiqe	Sat Jul 15 10:07:22 +0000 2023

3.1.1 *Cleaning*

Hal pertama yang dilakukan pada proses *preprocessing* ialah melakukan *cleaning* yaitu menghapus atau menghilangkan tanda baca, angka, URL, *mention* (@), dan *hashtag* (#) pada data *scraping* dengan menggunakan *library tm (Text Mining).* Hasil data *cleaning* dapat dilihat pada Tabel 2. Setelah dilakukan rangkaian dari proses *preprocessing*, didapatkan hasil data yang dapat dilihat pada Tabel 3.

Tabel 2 Contoh Data Hasil *Cleaning*

No.	*Before Cleaning*	*After Cleaning*
1	@jenjensante sampai ada twit sejahat itu loh dan semua twit bully lainnya mayoritas dari solo standnya kpop	sampai ada twit sejahat itu loh dan semua twit bully lainnya mayoritas dari solo standnya kpop
2	Kalah di bully Menang di kira curang, pake vpn lah, minta bantuan fandom lain lah (one person vs kpop cenah) di kira fandom kpop pro ke exol semua apa, mereka juga punya bias nya buat di vote elahh.. Big Love buat EXO + L aja❤ EXO VS EXOL sebenernya mah lu mah gak di ajak	Kalah di bully Menang di kira curang pake vpn lah minta bantuan fandom lain lah one person vs kpop cenah di kira fandom kpop pro ke exol semua apa mereka juga punya bias nya buat di vote elahh Big Love buat EXO L aja EXO VS EXOL sebenernya mah lu mah gak di ajak
3	Kpop itu ibarat lu boleh bully gua, boleh ngejatuhin gue, tp lu jangan pernah mengusik obat kebahagiaan gue	pop itu ibarat lu boleh bully gua boleh ngejatuhin gue tp lu jangan pernah mengusik obat kebahagiaan gue
4	tp emang baek kooyoung di bully abis abisan, gak cuma sama anak bp dan zb1 tapi juga sama idol idol kpop lain. salah sendiri sih..... jahat banget pas ngomong😭	tp emang baek kooyoung di bully abis abisan gak cuma sama anak bp dan zb tapi juga sama idol idol kpop lain salah sendiri sih jahat banget pas ngomong
5	@RuNext_base ga punya adab emg kebiasaan tukang bully disekolah, dibawa ke kpop	ga punya adab emg kebiasaan tukang bully disekolah dibawa ke kpop

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

3.1.2 *Case Folding*

Case folding adalah proses mengubah semua huruf dalam teks memiliki karakter huruf yang senada yakni huruf kecil keseragaman dan konsistensi. Tujuan utamanya adalah untuk memudahkan analisis dan pemrosesan teks lebih lanjut, menghindari masalah yang timbul akibat variasi kasus dalam data, dan memungkinkan pencocokan dan perbandingan teks secara lebih efisien.

3.1.3 *Tokenizing*

Pada tahap pemrosesan ini teks akan dipecah menjadi unit-unit yang lebih kecil dan hasil dari proses *tokenizing* merupakan tiap kata yang ada pada kalimat atau bisa disebut dengan *token*. *Token* biasanya berupa kata-kata, simbol, angka, atau bagian-bagian penting lainnya dalam teks. Paket-paket seperti *tm* (*Text Mining*) atau *tokenizers* di R menyediakan fungsi-fungsi untuk melakukan *tokenizing* teks dengan berbagai metode, termasuk *tokenisasi* kata per kata, *tokenizing* berdasarkan n-gram (grup kata), atau *tokenizing* berdasarkan kalimat.

3.1.4 *Stopword Removal*

Tahap selanjutnya setelah melakukan proses *tokenizing* ialah melakukan *stopword removal*. *Stopword removal* sendiri adalah proses untuk menghapus kata-kata yang dianggap memiliki sedikit atau bahkan tidak ada kontribusi untuk pemahaman makna suatu teks. Kata-kata semacam ini biasanya merupakan kata-kata umum seperti "dan," "atau," "saya," "di," dan lain-lain, yang sering muncul di hampir semua teks tetapi tidak memberikan informasi yang bermakna. Pada tahap ini menggunakan *library* seperti *tidytext* dan *tidyverse*, dengan hasil akhirnya berupa teks yang lebih relevan dan fokus pada kata-kata penting dalam analisis teks, tanpa terpengaruh oleh kata-kata umum yang tidak memberikan makna khusus.

3.1.5 *Stemming*

Tahapan terakhir dalam proses *text preprocessing* adalah *stemming*. Pada tahap ini dilakukan proses penghilangan awalan atau akhiran pada kata dalam teks untuk mengembalikan kata ke bentuk dasarnya. Tujuan dari *stemming* adalah untuk mengurangi kata-kata yang berbeda tetapi memiliki akar kata yang sama menjadi bentuk yang sama sehingga mempermudah analisis teks. Pada penelitian ini, untuk melakukan proses *stemming* data pada Rstudio akan digunakan *library* yaitu *katadasaR* dan *SnowballC*.

Tabel 3 Contoh Data Hasil *Preprocessing*

No.	*Stemming*
1	['sampai', 'ada', 'twit', 'sejahat', 'semua', 'twit', 'bully', 'lainnya', 'mayoritas','solo', 'standnya', 'solo']
2	['kalah', 'bully', 'menang', 'kira', 'curang', 'pake', 'vpn', 'lah', 'minta', 'bantuan', 'fandom', 'lain', 'lah', 'one', 'person', 'vs', 'kpop', 'cenah', 'kira', 'fandom', 'kpop', 'pro', 'exol', 'semua', 'apa', 'mereka', 'juga', 'punya', 'bias', 'buat', 'vote', 'elahh', 'big', 'love', 'buat', 'exo', 'l', 'exo', 'vs', 'exol', 'sebenernya', 'mah', 'mah', 'ajak']
3	['kpop', 'ibarat', 'boleh', 'bully', 'gua', 'boleh', 'ngejatuhin', 'gue', 'tp', 'jangan', 'pernah', 'mengusik', 'obat', 'kebahagiaan', 'gue']
4	['tp', 'emang', 'baek', 'kooyoung', 'bully', 'abis', 'abisan', 'cuma', 'sama', 'anak', 'bp', 'zb', 'tapi', 'juga', 'sama', 'idol', 'idol', 'kpop', 'lain', 'salah', 'sendiri', 'jahat', 'banget', 'pas', 'ngomong']
5	['punya', 'adab', 'emg', 'kebiasaan', 'tukang', 'bully', 'disekolah', 'dibawa', 'kpop']

3.2 *Term Frequency*

Term Frequency merupakan metode paling sederhana dari tahapan pembobotan kata. pada tahapan ini dilakukan proses untuk mengukur seberapa sering sebuah *term* (kata) muncul dalam sebuah dokumen atau teks. Secara sederhana, *term frequency* adalah jumlah kemunculan suatu

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

kata tertentu dibagi dengan total jumlah kata dalam dokumen tersebut. Dokumen pada penelitian ini adalah data dari *tweets* sedangkan *term* adalah kata yang ada pada *tweets*. Setelah melakukan program *term frequency*, akan didapatkan hasil dari proses frekuensi data yang dapat dilihat pada Tabel 4. Bisa dilihat pada tabel tersebut, hasil dari proses *term frequensi* bahwa untuk term "*kpop*" memiliki total frekuensi sebesar 775, *term* "*bully*" memiliki nilai total frekuensi sebesar 723, "*fans*" sebesar 184, dan seterusnya.

Tabel 4 Contoh Data Hasil Proses *Term* Frekuensi

No.	*Term*	Frekuensi
1	"kpop"	775
2	"bully"	723
3	"fans"	184
4	"sama"	179
5	"ada"	169

3.3 *Word Cloud*

Tahapan yang selanjutnya dilakukan ialah tahapan *word cloud*. Pada tahapan proses ini, akan dilakukan representasi visual dari sekumpulan kata yang disajikan dalam bentuk gambar. Pada *Word Cloud*, ukuran dari setiap kata menunjukkan frekuensi kemunculannya dalam teks atau data yang dianalisis. Kata-kata yang muncul lebih sering akan ditampilkan dengan ukuran yang lebih besar dan lebih menonjol dalam *Word Cloud*. Pada tahap ini, dilakukan proses *word cloud* dengan menggunakan *library wordcloud2* untuk membuat dan *RcolorBrewer* untuk memilih palet warna. Pada tahap ini akan dilakukan pemanggilan atau penampilan *word cloud* dari *term* yang memiliki frekuensi lebih dari 30. Seperti yang ditampilkan pada Gambar 2, didapatkan bahwa *term* "*bully*" dan "*kpop*" merupakan kata yang paling sering dibicarakan para pengguna Twitter tentang topik *cyberbullying* pada *kpop*.

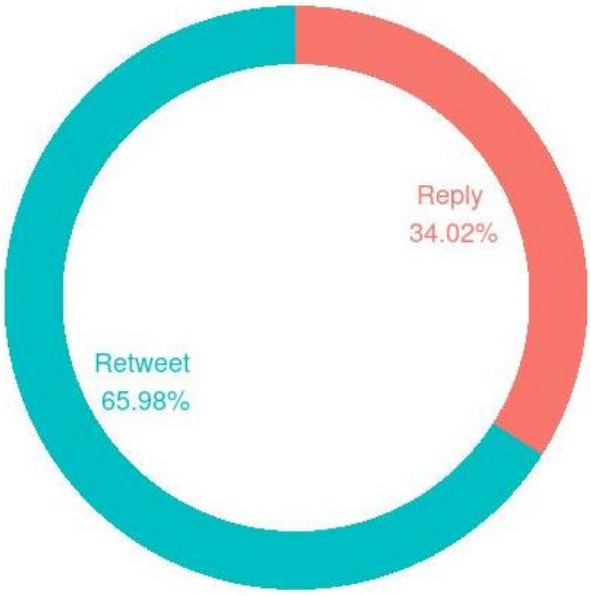
Gambar 2 Hasil *Word Cloud*

3.4 *Social Network Analysis* (SNA)

SNA, atau juga dikenal sebagai Analisis Jaringan Sosial, adalah suatu metode ilmiah yang bertujuan untuk memahami dan menganalisis hubungan sosial antara individu, kelompok, atau organisasi. Metode ini menggambarkan dan mengukur interaksi dan koneksi antara entitas dalam suatu jaringan. Metode ini dapat digunakan untuk mengidentifikasi *node* yang penting dalam

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

suatu jaringan jejaring sosial seperti Twitter, dalam studi kasus *cyberbullying kpop*. Dalam penelitian ini, interaksi antar *node* dalam jaringan tersebut dikelompokkan menjadi dua jenis yakni *reply* (balasan) dan *retweet*. *Node-node* yang terlibat dalam interaksi tersebut adalah akun-akun pengguna yang aktif dalam mengirimkan *tweet* tentang kasus *cyberbullying kpop* yang terjadi di Twitter. Program yang digunakan untuk melihat persentase *retweet* dan *reply* menggunakan bantuan sebuah *library* yaitu *ggplot2* dan *tidyverse* yang nantinya akan menampilkan hasil persentase yang dapat dilihat pada Gambar 3.



Gambar 3 Hasil Persentase *Retweet* dan *Reply*

Seperti yang terlihat pada Gambar 3, bahwa persentase tertinggi adalah jenis interaksi dari *retweet* yaitu sebesar 65,98%, dan interaksi dari reply memiliki persentase yang sebesar 34,02%. Pada penelitian ini juga menggunakan data dalam bentuk *edge list*. *Edge list* merupakan representasi grafik yang paling sederhana dan umum digunakan yang terdiri atas dua kolom yang mencatat setiap sambungan dalam grafik. Setiap baris dalam daftar tersebut mencantumkan dua simpul (*node*) yang terhubung oleh suatu "sisi" (*edge*).

Pada tahapan ini, *software* Gephi-0.9.5 digunakan untuk melakukan analisa pada metode SNA dengan menggunakan jenis graf yaitu *Undirected graph*. Dengan menggunakan jenis graf tersebut maka arah dari interaksi *node* tidak diperhatikan. Setelahnya, akan didapatkan nilai *node* sebanyak 778 dengan *edges* sebanyak 1051. Artinya, dalam data *cyberbullying kpop* yang dianalisa, terdapat 778 akun dan terdapat 1051 interaksi yang terjadi.

Langkah selanjutnya ialah mencari nilai *modularity* guna mengukur sejauh mana suatu grafik terbagi menjadi kelompok-kelompok (*communities*) yang lebih padat secara internal dan lebih jarang terhubung satu sama lain. Pada penelitian ini nilai *modularity* yang didapatkan adalah 0,938 dan didapat juga jumlah *communities* sebanyak 192. Selanjutnya ialah mencari nilai *centrality* yang nantinya akan digunakan untuk mengidentifikasikan aktor yang terpenting.

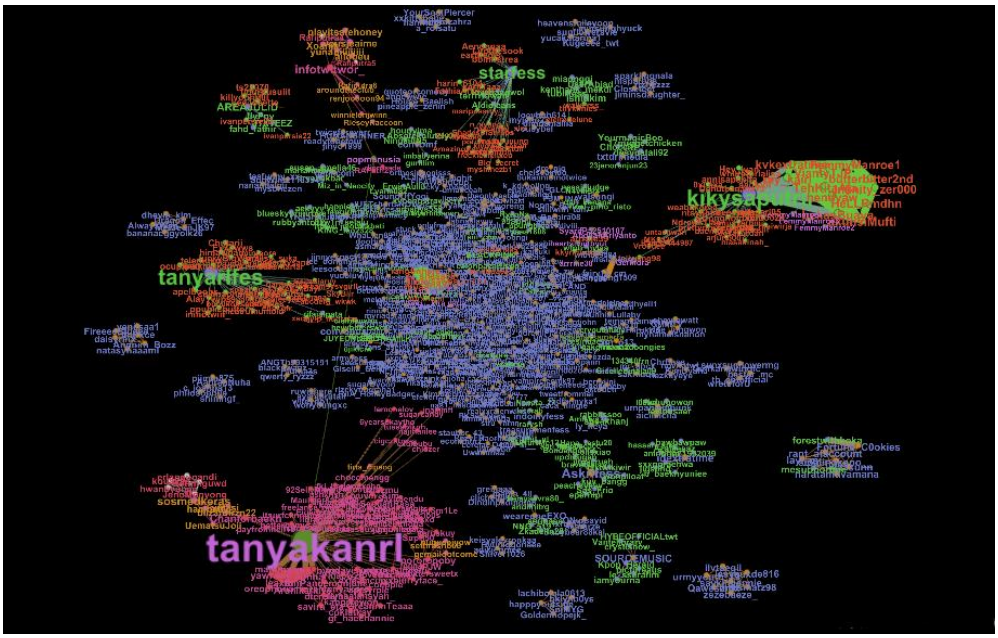
3.4.1 *Degree Centrality*

Degree centrality digunakan untuk mengukur seberapa banyak simpul dalam grafik memiliki koneksi langsung. Sederhananya, metrik ini menghitung jumlah *edge* yang terhubung ke simpul tertentu. Simpul dengan *degree centrality* tinggi berperan penting dalam menyebarkan informasi dan interaksi dalam jaringan. Pada Tabel 5 menampilkan 5 *node* yang mempunyai nilai *degree centrality* tertinggi.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Tabel 5 Data *Node* dengan *Degree Centrality* Teratas

Username	*Degree Centrality*	Deskripsi Akun
tanyakanrl	80	Sebuah akun Menfess bot yang diciptakan untuk berbagi informasi atau cerita seputar topik Reinforcement Learning (RL), memiliki 1,2 juta pengikut.
kikysaputrii	38	Seorang komedian yang merupakan salah satu finalis di acara Stand Up Comedy musim keempat, memiliki 167,1 ribu pengikut.
tanyarlfes	37	Sebuah akun Menfess bot yang diciptakan untuk berbagi informasi atau cerita seputar topik Reinforcement Learning (RL), memiliki 862,3 ribu pengikut.
starfess	25	Sebuah akun Menfess bot yang diciptakan untuk berbagi info seputar figur publik di seluruh dunia, memiliki 799,4 ribu pengikut.
Askrlfess	14	Sebuah akun automenfess yang membahas tentang hal apapun, RL, lucu-lucuan dll, memiliki 806,5 ribu pengikut.



Gambar 4 Data Visualisasi *Degree Centrality*

Dapat dilihat dari data pada Tabel 5, bahwa *node* dengan nilai *degree centrality* terbanyak adalah "*tanyakanrl*" dengan nilai sebanyak 80 yang merupakan Sebuah akun Menfess bot yang diciptakan untuk berbagi informasi atau cerita seputar topik *Reinforcement Learning* (RL), posisi kedua yaitu akun "*kikysaputrii*" dengan nilai *degree centrality* yaitu 38 merupakan seorang komedian yang juga menjadi salah satu *finalis* pada ajang *Stand Up Comedy* musim keempat, dan pada posisi ketiga yaitu akun "*tanyarlfes*" dengan nilai yaitu 37 yang juga merupakan sebuah akun Menfess bot yang diciptakan untuk berbagi informasi atau cerita seputar topik *Reinforcement Learning* (RL). Dengan visualisasi *degree centrality* yang dapat dilihat pada Gambar 4.

3.4.2 *Betweenness Centrality*

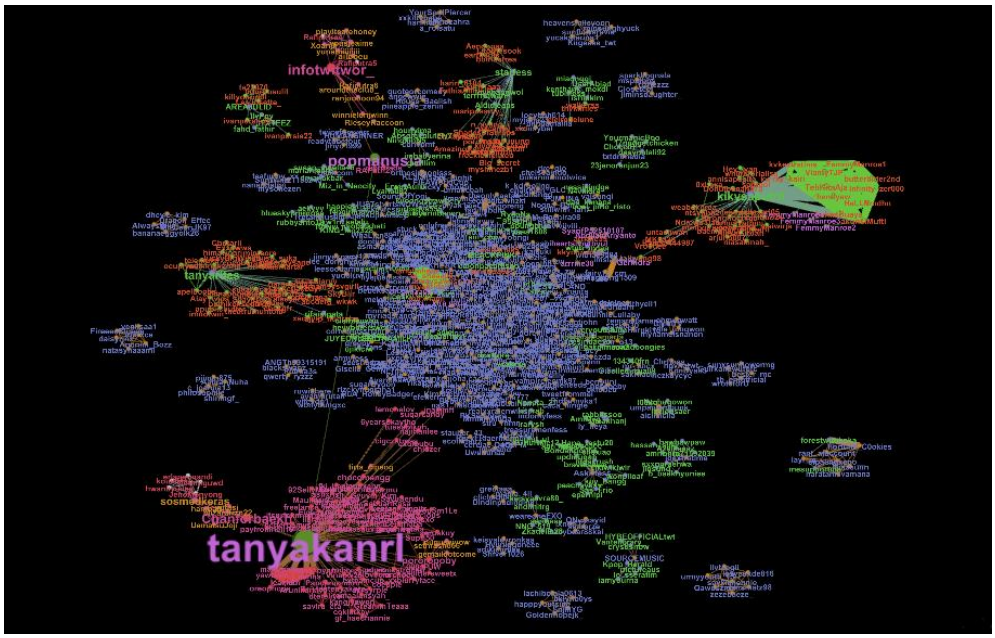
Betweenness centrality mengukur sejauh mana suatu simpul berada di antara jalur terpendek antara pasangan simpul lain dalam grafik. Simpul dengan *betweenness centrality* tinggi berperan sebagai penghubung utama dalam jaringan, menghubungkan kelompok-kelompok yang

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

berbeda, dan mengontrol aliran informasi. Pada Tabel 6, menampilkan 5 *node* yang mempunyai nilai *betweenness centrality* paling tinggi.

Tabel 6 Data *Node* dengan *Betweenness Centrality* Teratas

Username	*Betweenness Centrality*	Deskripsi Akun
tanyakanrl	5594.2	Sebuah akun Menfess bot yang diciptakan untuk berbagi informasi atau cerita seputar topik Reinforcement Learning (RL), memiliki 1,2 juta pengikut.
popmanusia	1437,0	Akun yang berisi konten-konten yang membahas isu-isu dan masalah-masalah terkini, memiliki 48 pengikutdan bergabung pada Juni 2016.
infotwitwor_	1208,0	Akun yang berisi konten-konten yang membahas isu-isu dan drama yang terjadi di Twitter, memiliki 693,9 ribu pengikut dan bergabung pada Februari 2020.
Chanforbaekh	990,0	Merupakan akun dari fandom EXO yang membahas berita betita terkini mengenai para personil grup EXO terutama Baekhyun dan Chanyeol.
kikysaputrii	747,0	Seorang komedian yang merupakan salah satu finalis di acara Stand Up Comedy musim keempat, memiliki 167,1 ribu pengikut.



Gambar 5 Data Visualisasi *Betweenness Centrality*

Dari Tabel 6, bisa dilihat bahwa akun "*tanyakanrl*" memiliki nilai *betweenness centrality* terbesar yaitu 5.594,2 merupakan sebuah akun *menfess* bot yang diciptakan untuk berbagi informasi atau cerita seputar topik *Reinforcement Learning* (RL) dan memiliki 1,2 juta pengikut. pada posisi dua terdapat akun "*popmanusia*" dengan nilai sebesar 1.437,0 merupakan akun yang berisi konten-konten yang membahas isu-isu dan masalah-masalah terkini, dan diikuti akun "*infotwitwor_*" dengan nilai *betweenness centrality* yaitu 1.208,0 merupakan akun yang berisi konten-konten yang membahas isu-isu dan drama yang terjadi di Twitter dan memiliki 693,9 ribu pengikut. Bisa diartikan bahwa "*tanyakanrl*" merupakan *node* yang paling efektif dalam menghubungkan *communities* yang terdapat pada data *cyberbullying kpop*. Pada Gambar 5 menampilkan visualisasi dari data *betweenness centrality*.

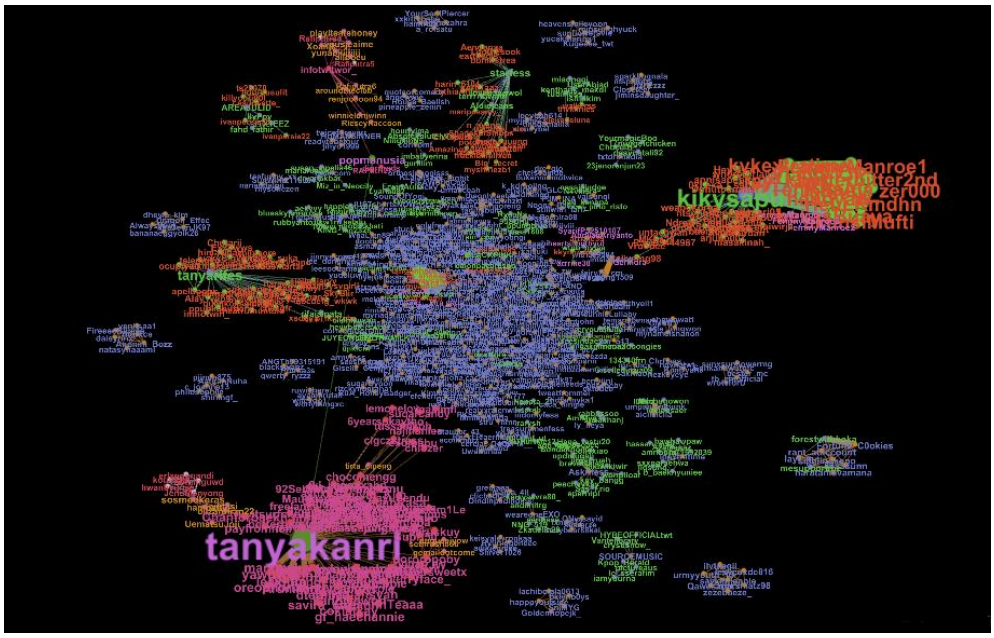
Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

3.4.3 *Eigenvector Centrality*

Eigenvector centrality menilai pentingnya suatu simpul berdasarkan seberapa banyak simpul-simpul lain yang terhubung dengannya. Simpul dengan *eigenvector centrality* tinggi cenderung terhubung dengan simpul-simpul penting lainnya, sehingga memiliki pengaruh yang besar dalam jaringan. Pada Tabel 7 menampilkan 5 *node* yang mempunyai nilai *eigenvector centrality* tertinggi.

Tabel 7 Data *Node* dengan *Eigenvecor Centrality* Teratas

Username	*Eigenvector Centrality*	Deskripsi Akun
tanyakanrl	1,0	Sebuah akun Menfess bot yang diciptakan untuk berbagi informasi atau cerita seputar topik Reinforcement Learning (RL), memiliki 1,2 juta pengikut.
kikysaputrii	0,549206	Seorang komedian yang merupakan salah satu finalis di acara Stand Up Comedy musim keempat, memiliki 167,1 ribu pengikut.
akunsiMufti	0,381676	Akun yang berisi aktifitas sehari-hari, memiliki 364 pengikut dan bergabung pada Agustus 2021.
bekasBuaya	0,381676	Akun yang berisi konten-konten yang membahas isu-isu terkini, memiliki 171 pengikut dan bergabung pada September 2021.
TehKitaAja	0,381676	Akun yang berisi konten-konten yang membahas isu-isu terkini, memiliki 983 pengikut dan bergabung pada September 2020.



Gambar 6 Data Visualisasi *Eigenvector Centrality*

Pada Tabel 7, diketahui bahwa nilai *eigenvector centrality* terbesar dipegang oleh sebuah akun bot *menfess* yaitu "*tanyakanrl*" dengan nilai sebesar 1,0, diikuti pada posisi kedua oleh akun seorang komedian dan juga seorang *stand up comedian* "*kikysaputrii*" yang mempunyai nilai *eigenvector centrality* yaitu 0,549206, lalu pada posisi dengan nilai *eigenvector centrality* sebesar 0,381676 yaitu "*akunsiMufti*" yang merupakan sebuah akun yang berisi aktivitas sehari-hari dengan 364 pengikut dan bergabung pada Agustus 2021. Jadi, dapat disimpulkan bahwa "*tanyakanrl*" merupakan *node* yang paling berpengaruh dalam penyebaran informasi mengenai

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

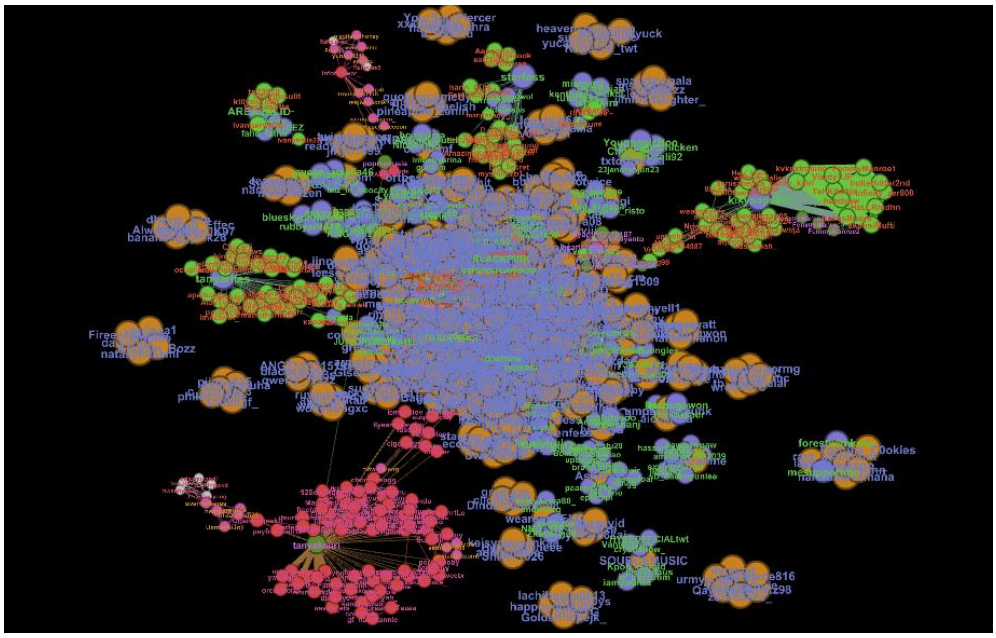
cyberbullying kpop. Berikut ini merupakan visualisasi dari hasil data *eigenvector centrality* yang ditampilkan pada Gambar 6.

3.4.4 *Closeness Centrality*

Closeness centrality mengukur seberapa cepat suatu simpul dapat mencapai simpul-simpul lain dalam jaringan melalui jalur terpendek. Simpul dengan *closeness centrality* tinggi memiliki akses yang tercepat ke informasi dan sumber daya dalam jaringan. Pada Tabel 8, menampilkan 5 *node* yang mempunyai nilai *closeness centrality* tertinggi. Bisa dilihat pada tabel tesebut, bahwa nilai *closeness centrality* dari 5 akun teratas memiliki nilai yang sama yaitu 1,0. Hal ini berarti bahwa kelima akun tersebut merupakan akun yang cepat dalam mengakses informasi terkait *cyberbullying kpop*. Pada Gambar 7 menampilkan visualisasi dari hasil data *closeness centrality*.

Tabel 8 Data *Node* dengan *Closeness Centrality* Teratas

Username	*Closeness Centrality*	Deskripsi Akun
layl_m	1,0	Akun yang berisi bahasan tentang masalah-masalah kpop terkini, memiliki 216 pengikut dan bergabung pada Mei 2015.
naratamavamana	1,0	Akun yang berisi konten-konten yang membahas isu-isu terkini, memiliki 445 pengikut dan bergabung pada Maret 2013.
Fortune_C0okies	1,0	Akun yang berisi konten-konten yang membahas isu-isu terkini.
rant_afaccount	1,0	Akun yang berisi konten-konten yang membahas isu-isu terkini.
sipalingkepo__	1,0	Akun penggemar BTS yang berisi konten tentang aktifitas group BTS.



Gambar 7 Data Visualisasi *Closeness Centrality*

3.5 Identifikasi Aktor

Berdasarkan hasil analisis, akun Twitter dengan *username* "*tanyakanrl*" merupakan aktor yang paling penting dalam jaringan data *cyberbullying kpop*. Akun ini memiliki nilai tertinggi pada

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

matriks *Degree Centrality*, *Betweenness Centrality*, dan *Eigenvector Centrality*, menandakan perannya dalam menyebarkan informasi dan interaksi dalam jaringan data *cyberbullying kpop*. Pada bulan Juli 2023, akun "*tanyakanrl*" yang terlihat pada Gambar 8, telah mencapai 1,2 juta pengikut dan termasuk dalam kategori *autobase* di Twitter. Akun *autobase* adalah akun otomatis yang digunakan untuk berdiskusi dan berbagi informasi antara basis penggemar atau kelompok tertentu melalui "*menfess*" (pengakuan anonim). Akun-akun *autobase* seperti ini aktif membahas berbagai topik, termasuk isu serius seperti *cyberbullying* di industri K-pop.

Gambar 8 Akun Twitter tanyakanrl

Keberadaan akun *autobase* seperti "*tanyakanrl*" memungkinkan informasi tentang isu *cyberbullying* di industri K-pop menyebar luas dengan cepat karena memiliki jumlah pengikut yang besar dan mencapai khalayak yang luas. Isu ini menjadi perhatian penting dan menjadi pembicaraan di kalangan komunitas penggemar K-pop berkat adanya akun *autobase* seperti "*tanyakanrl*". Penelitian ini juga mengidentifikasi peran media sosial dalam membentuk narasi *cyberbullying* di kalangan fandom K-pop (Cho & Lee, 2022).

4. KESIMPULAN

Dari penelitian "Analisa Jejaring Sosial Terhadap Fenomena *Cyberbullying Fandom* K-Pop Pada Sosial Media Twitter" dapat diambil beberapa kesimpulan. Pertama, penelitian ini mengumpulkan data sebanyak 739 *tweet* dengan persentase *retweet* sebesar 65,98% dan *reply* sebesar 34,02%. Dari hasil analisis *word cloud*, kata-kata yang sering digunakan dalam topik *cyberbullying kpop* adalah "*kpop*," "*bully*," dan "*fans*."

Kedua, berdasarkan hasil analisis *centrality*, ditemukan bahwa *node* yang paling penting dalam jaringan data *cyberbullying kpop* adalah akun Twitter dengan *username* "*tanyakanrl*." Akun ini memiliki peran yang signifikan dalam menyebarkan informasi agar tidak terjadi perundungan di kalangan penggemar kpop. Serta nteraksi dalam topik *cyberbullying kpop* berdasarkan nilai *degree centrality*, *betweenness centrality*, *eigenvector centrality*, dan *closeness centrality*.

Ketiga, akun "*tanyakanrl*" termasuk dalam kategori *autobase* di Twitter, yang merupakan akun otomatis digunakan untuk berdiskusi dan berbagi informasi antara basis penggemar atau kelompok tertentu melalui "*menfess*" (pengakuan anonim). Dengan demikian, penelitian ini memberikan pemahaman tentang pola interaksi dan peran penting akun-akun tertentu dalam topik *cyberbullying kpop* di platform media sosial Twitter, khususnya peran yang dimainkan oleh akun "*tanyakanrl*" sebagai aktor utama dalam jaringan tersebut. Penyebaran pesan-pesan *cyberbullying* dalam jaringan fandom K-pop di Twitter dapat dianalisis lebih lanjut untuk memahami dinamika dan dampak psikologis yang ditimbulkannya (Park & Kim, 2023). Terakhir,

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

pentingnya penelitian lebih lanjut tentang dampak psikologis dari *cyberbullying* dalam fandom K-pop di Twitter telah diakui (Lim & Choi, 2024).

DAFTAR PUSTAKA

Akbar, Z. I. (2021, April 23). *Apa itu Text Mining?* BINUS Higher Education. https://sis.binus.ac.id/2021/04/23/apa-itu-text-mining/

Bader, D. A., Kintali, S., Madduri, K., & Mihail, M. (2007). Approximating Betweenness Centrality. In *Algorithms and Models for the Web-Graph: Vol. 4863 LNCS* (pp. 124–137). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77004-6_10

Brandes, U. (2001). A faster algorithm for betweenness centrality*. *The Journal of Mathematical Sociology*, *25*(2), 163–177. https://doi.org/10.1080/0022250X.2001.9990249

Cho, Y., & Lee, J. (2022). The Role of Social Media in Shaping the Narrative of Cyberbullying in K-Pop Fandoms. *Social Media + Society*, *8*(1).

Donegan, R. (2012). Bullying and Cyberbullying : History, Statistics, Law, Prevention and Analysis. *The Elon Journal of Undergraduate Research in Communications*, *3*(1), 33–42. https://api.semanticscholar.org/CorpusID:16065546

Gradinger, P., Strohmeier, D., & Spiel, C. (2010). Definition and Measurement of Cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *4*(2). https://cyberpsychology.eu/article/view/4235/3280

Irawan, D. (2018). *Psikolog: Cyberbullying Bisa Membuat Korban Jadi Depresi*. Health Liputan6.Com. https://www.liputan6.com/health/read/3304433/psikolog-cyberbullying-bisa-membuat-korban-jadi-depresi

Kim, H. J., & Park, S. H. (2021). Sentiment Analysis of Tweets Related to K-Pop Fandoms: Focusing on Cyberbullying Incidents. *International Journal of Data Science and Analytics*, *11*(3), 287–299.

Lee, S., & Jang, J. (2020). Exploring the Dynamics of Cyberbullying in K-Pop Fandom Communities on Twitter. *Journal of Cyberpsychology, Behavior, and Social Networking*, *23*(5), 315–322.

Lim, J. H., & Choi, D. (2024). A Study on the Psychological Impact of Cyberbullying within K-Pop Fandoms on Twitter. *Journal of Media Psychology*, *36*(2), 223–237.

Park, M., & Kim, S. (2023). Analyzing the Spread of Cyberbullying Messages in K-Pop Fandom Networks on Twitter. *Journal of Computational Social Science*, *6*(2), 145–160.

Tineges, R., & Davita, A. W. (2021). *Tahapan Text Preprocessing dalam Teknik Pengolahan Data*. DQLab. https://dqlab.id/tahapan-text-preprocessing-dalam-teknik-pengolahan-data

Tionardus, M., & Setuningsih, N. (2022, August 18). *Arti Fandom di Kpop*. Kompas.Com. https://entertainment.kompas.com/read/2022/08/18/152828066/arti-fandom-di-kpop?page=all

Tsvetovat, M., & Kouznetsov, A. (2011). *Social Network Analysis for Startups: Finding Connections on the Social Web*. O'Reilly Media, Inc. https://books.google.co.in/books?id=Tn-L5WoCeygC&printsec=frontcover#v=onepage&q&f=false

Yuniar, E., Utsalinah, D. S., & Wahyuningsih, D. (2022). Implementasi Scrapping Data Untuk Sentiment Analysis Pengguna Dompet Digital dengan Menggunakan Algoritma Machine Learning. *Jurnal Janitra Informatika Dan Sistem Informasi*, *2*(1), 35–42. https://doi.org/10.25008/JANITRA.V2I1.145

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024, Pp. 94 – 104
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Analisis Keamanan Data Pelanggan dalam Menghadapi Tantangan Penggunaan *Marketplace*

Rizki Dewantara [(1)*], Rauhulloh Ayatulloh Khomeini Noor Bintang [(2)], Rahmadhan Gatra [(3)]
[1] Sains Data, Fakultas Sains Teknologi dan Kesehatan, Institut Teknologi Bisnis Dan Kesehatan Bhakti Putra Bangsa Indonesia, Purworejo
[2] Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Karanganyar, Karanganyar
[3] UPT. PTIPD UIN Sunan Kalijaga, Yogyakarta
e-mail : dewantararizki@ibisa.ac.id, rauhulloh.bintang@umuka.ac.id, rahmadhan.gatra@uin-suka.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 7 Desember 2023, direvisi 7 April 2024, diterima 15 April 2024, dan dipublikasikan 25 Mei 2024.

Abstract

The advent of the digital economy makes commerce more accessible to everyone. Example: an online marketplace app that simplifies buying and selling. The marketplace app's useful features and ease of use attract many users. The marketplace app's functionality and convenience have been enhanced to meet consumer expectations and prioritize consumer data protection. This study investigates how customers protect their data when shopping online and utilizing marketplace apps. Environmental and social influences, personal data security facilities, the goal of utilizing the marketplace, and awareness of customer data security when using the marketplace application were asked of 70 random sample participants. The questionnaires had 16 Guttman scale questions. According to the report, 81.42% of customers trust the marketplace app to protect their data. Likewise, 88.57% of customers strongly believe that the marketplace application they use secures their personal information, indicating that this is related to their marketplace service needs.

Keywords: Customer, Data Security, Marketplace, Capability, Digital Economy, User Satisfaction

Abstrak

Munculnya ekonomi digital membuat perdagangan menjadi lebih mudah bagi semua orang. Contoh: aplikasi *marketplace* online yang mempermudah jual beli. Fitur-fitur berguna dan kemudahan penggunaan aplikasi *marketplace* menarik banyak pengguna. Fungsionalitas dan kenyamanan aplikasi *marketplace* telah ditingkatkan untuk memenuhi harapan konsumen dan memprioritaskan perlindungan *data* konsumen. Studi ini menyelidiki seberapa sukses pelanggan melindungi *data* pribadi mereka saat berbelanja online dan menggunakan aplikasi *marketplace*. Pengaruh lingkungan dan sosial, fasilitas keamanan *data* pribadi, tujuan pemanfaatan *marketplace*, dan kesadaran akan keamanan *data* pelanggan saat menggunakan aplikasi *marketplace* ditanyakan kepada 70 peserta sampel secara acak. Kuesioner memiliki 16 pertanyaan skala *Guttman*. Menurut laporan tersebut, 81,42% pelanggan memercayai aplikasi *marketplace* untuk melindungi *data* mereka. Begitupun dengan 88,57% pelanggan sangat yakin bahwa aplikasi *marketplace* yang mereka gunakan mengamankan informasi pribadi mereka, hal ini menunjukkan bahwa hal ini berkaitan dengan kebutuhan layanan *marketplace* mereka.

Kata Kunci: Pelanggan, Keamanan Data, Pasar Daring, Kemampuan, Ekonomi Digital, Kepuasan Pengguna

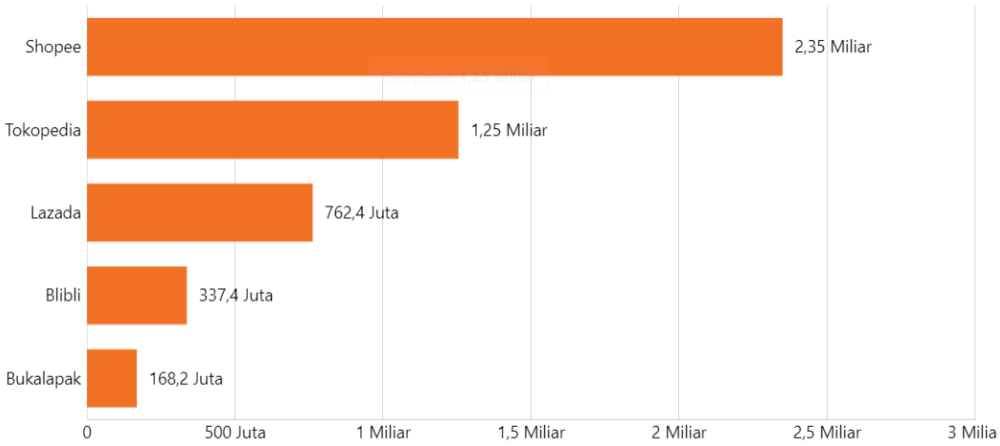
1. PENDAHULUAN

Suatu organisasi atau perusahaan tidak akan mempercayakan datanya kepada penyedia penyimpanan tanpa terlebih dahulu memastikan bahwa penyedia tersebut memperhatikan perlindungan data dengan serius (Ula, 2019). Tidak akan pernah ada peraturan final yang mengatur penanganan data pribadi secara adil. Dalam mengikuti perubahan zaman, peraturan

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

tersebut memerlukan pembaruan dan amandemen secara berkala, serupa dengan undang-undang perlindungan konsumen atau lingkungan hidup (Hoofnagle et al., 2019). Jaringan transmisi yang tidak aman merupakan sumber data konsumen yang berpotensi jatuh ke tangan pihak-pihak yang tidak bertanggung jawab dalam jaringan data (Putra et al., 2019). Ketika privasi data menjadi perhatian publik yang penting, pembelajaran kooperatif telah muncul sebagai garda depan penelitian dalam upaya memfasilitasi pengembangan model pembelajaran kolaboratif di berbagai organisasi yang terlibat dalam privasi data (Li et al., 2023). Namun, terdapat juga data berukuran besar yang menimbulkan risiko dan tantangan; di antaranya adalah pertanyaan penting tentang privasi pengguna (Price & Cohen, 2019). Ada peluang baru untuk meningkatkan kualitas layanan untuk aplikasi baru melalui berbagi data, berkat pertumbuhan eksponensial dalam volume data yang dihasilkan oleh perangkat yang terhubung dalam paradigma industri Internet of Things (IoT). Penyedia data mempunyai tantangan yang signifikan ketika mencoba mentransfer data menggunakan jaringan nirkabel karena masalah keamanan dan privasi (seperti kebocoran data). Selain menyebabkan penyedia layanan merugi, dampak kebocoran data pribadi bisa lebih parah lagi (Lu et al., 2020).

Seiring berkembangnya sektor digital, semakin banyak data pribadi yang digunakan. Aplikasi *online*, seperti Teknologi Finansial dan transaksi *online*, mengharuskan pengguna untuk mengungkapkan data pribadi agar dapat mengakses dan memanfaatkan layanan ini. Meskipun demikian, terdapat bahaya kebocoran dan kerugian sosial yang terkait dengan penggunaan data pribadi (Kesuma et al., 2021). Kekhawatiran terhadap keamanan data, yang berkaitan dengan perlindungan informasi pribadi milik semua komunitas, berpotensi berdampak signifikan terhadap kepercayaan masyarakat terhadap *e-government*. Meski demikian, pengamanan data pribadi saat ini tidak diatur oleh peraturan perundang-undangan di Indonesia. Karena negara mempunyai kewajiban untuk menjamin privasi informasi warga negaranya, maka penting bagi pemerintah untuk menetapkan kerangka legislatif untuk menjamin keamanan sistem elektroniknya (Iswandari, 2021). Hasil pengujian menunjukkan bahwa harga, kualitas, dan promosi berpengaruh sebesar 69,4% terhadap variabel keputusan pembelian, sedangkan faktor lain atau variabel independen seperti volume penjualan, margin distribusi, kualitas pelayanan, dan lain-lain memberikan pengaruh sebesar 30,6% (Prilano et al., 2020). Pelanggan lebih cenderung berbelanja di retail *online* Bukalapak setelah menggunakan OCR. Niat membeli *online* di kalangan Bukalapak masing-masing dipengaruhi secara positif oleh kepercayaan pelanggan dan OCR (Mulyati & Gesitera, 2020). Instagram, Facebook, dan Whatsapp bukanlah platform pemasaran digital terbaik untuk produk perusahaan. Teknik periklanan yang tidak konsisten dan promosi yang tidak tepat sasaran adalah dua masalah yang mengganggu upaya pemasaran digital bisnis.



Gambar 1 Kunjungan Situs *E-Commerce Marketplace*

Sepeti yang digambarkan pada Gambar 1, selama Januari hingga Desember 2023, lalu lintas situs web kumulatif Shopee mencapai 2,3 juta, yang lebih dari cukup untuk memuaskan para

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

pengiklan. Selama periode waktu yang sama, Tokopedia memiliki sekitar 1,2 juta kunjungan, sedangkan Lazada menerima 762,4 juta. Sementara Bukalapak menarik 168,2 juta pengunjung dan BliBli 337,4 juta (Ahdiat, 2024). Untuk menjangkau audiens yang lebih besar dengan biaya lebih sedikit dibandingkan sebelumnya, pemasaran digital adalah cara yang tepat (Rachmadewi et al., 2021). *E-satisfaction* menunjukkan bahwa dipengaruhi secara positif dan signifikan oleh kualitas informasi. Kepercayaan memediasi hubungan yang menguntungkan dan signifikan secara statistik antara kepuasan elektronik dan kegunaan yang dirasakan (Prajoko et al., 2022) dengan meningkatnya tingkat persaingan di *online marketplace*, loyalitas klien memainkan peran penting bagi bisnis (Karina, 2019).

Berdasarkan penelitian-penelitian sebelumnya, penelitian ini bertujuan untuk menyelidiki keamanan data pribadi selama penggunaan aplikasi. Untuk mengetahui pengalaman pelanggan terhadap aplikasi *marketplace*, penelitian ini menggunakan strategi penyebaran kuesioner. Berdasarkan asumsi bahwa melindungi informasi pribadi pelanggan akan diutamakan dibandingkan semua pertimbangan lain saat merancang aplikasi *marketplace*.

2. METODE PENELITIAN

2.1 Keamanan Data

Dengan menjaga keamanan data, hal ini dapat mengurangi kemungkinan bahaya dan ancaman dunia maya, menjaga operasional tetap berjalan lancar, dan membuka *marketplace* baru untuk investasi. Solusi teknis harus diterapkan untuk memenuhi tuntutan keamanan dan eksploitasi *data* pada saat yang bersamaan (Kaissis et al., 2020). Meskipun demikian, seiring dengan meningkatnya ketergantungan terhadap teknologi informasi dalam dunia usaha, kemungkinan terjadinya kerugian teknis dan non-teknis, serta ancaman dan pelanggaran terhadap kebijakan keamanan informasi organisasi juga meningkat. Pemeliharaan *data* yang tepat sangat penting bagi setiap perusahaan karena *data* adalah aset paling berharga bagi perusahaan mana pun (Putri et al., 2020).

2.2 *Marketplace*

Istilah *marketplace* mengacu pada bisnis apa pun yang menggunakan teknologi elektronik untuk tujuan pembelian, penjualan, dan promosi produk dan layanan (Hasanah, 2019). Dengan bantuan pemasaran media sosial, bisnis dapat lebih memahami keinginan dan kebutuhan pelanggan serta menyesuaikan produk dan layanan mereka untuk memenuhi permintaan tersebut. Dalam pemasaran, brand *awareness* mengacu pada sejauh mana konsumen memiliki pengetahuan dan kenyamanan menggunakan produk atau layanan yang menyandang nama dan merek dagang tertentu (Adrian & Mulyandi, 2021).

2.3 Kepuasan Pengguna

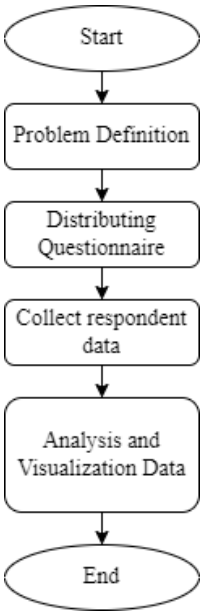
Era digital berkembang dengan sangat cepat dan tanpa peringatan. Banyaknya teknologi inovatif yang meningkatkan dan memudahkan kehidupan manusia lahir di era digital (Yang & Sihotang, 2023). Komponen kunci keberhasilan suatu aplikasi adalah tingkat kebahagiaan yang dirasakan penggunanya setelah menggunakannya (Ariska & Amelia, 2021). Memproses sentimen konsumen memungkinkan perolehan informasi berharga yang dapat dimanfaatkan baik oleh toko maupun konsumen lainnya (Sari & Wibowo, 2019), atribut produk dan nilai promosi terhadap niat beli konsumen (Suhatman et al., 2020), Variabel Kualitas Pelayanan, Produk, dan Harga semuanya berpengaruh positif terhadap variabel Minat Beli, baik sendiri maupun gabungan. Pelanggan lebih cenderung melakukan pembelian ketika mereka puas dengan layanan, produk, dan harga (Bakti & Septijantini Alie, 2020).

2.4 Alur Penelitian

Metode distribusi survei digunakan dalam penyelidikan ini. Penulis melakukan survei pengguna Tokopedia dan Shopee menggunakan kuesioner untuk mengetahui bagaimana kekhawatiran

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

pelanggan terhadap keamanan *data* memengaruhi minat mereka berbelanja di platform tersebut. Melalui Google Form, 70 peserta akan dikirimi survei online untuk diisi sesuai dengan teknik sampel acak. Kuesioner yang digunakan dalam penelitian ini adalah kuesioner terbuka. Kuesioner berisi enam belas pertanyaan yang menggunakan skala *Guttman*, memberikan pilihan jawaban ya/benar dan tidak/salah. Alur penelitian yang dilakukan digambarkan pada Gambar 2.



Gambar 2 Alur Penelitian

Tahap kajian dimulai dengan mendefinisikan masalah analisis belanja di *marketplace* yang berkaitan dengan keamanan data klien. Hal berikutnya yang dilakukan adalah membuat survei dan mengirimkannya kepada masyarakat untuk mengetahui bagaimana perasaan mereka terhadap kebijakan *marketplace* mengenai perlindungan data konsumen. Visualisasi data kemudian dilakukan setelah analisis *data* responden yang diperoleh dari kuesioner.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Responden

Tabel 1 Jumlah Responden Berdasarkan Gender

Gender	*Number of Respondents*	*Percentage* (%)
Laki-laki	32	45,7
Perempuan	38	54,3
Total	**70**	**100**

Berdasarkan distribusi gender responden yang ditampilkan pada Tabel 1, 45,7% dari 70 responden adalah laki-laki, sedangkan 54,3% adalah perempuan. Untuk mengetahui kebenaran setiap item instrumen, dilakukan uji validitas. Data yang digunakan adalah 70 orang sebagai sampel dari kelompok masyarakat yang merupakan pengguna *marketplace* yang terdiri berdasarkan dari umur 15 hingga 40 tahun dari berbagai pekerjaan seperti mahasiswa, karyawan, dan masyarakat umum sebagai uji coba instrumen. Menurut Singarimbun dan Effendi (1995), yang mengatakan bahwa jumlah minimum responden untuk uji coba kuesioner adalah minimal 30, maka jumlah sampel yang digunakan adalah 70 orang (Singarimbun & Effendi, 1995). Berdasarkan Tabel 2, kita dapat menentukan bahwa persentase pelanggan yang menggunakan *marketplace Shopee* adalah 54,4%, persentase pelanggan yang menggunakan *marketplace Tokopedia* adalah sekitar 17,1%, persentase pelanggan yang menggunakan *marketplace Lazada* adalah sekitar 12,8%, persentase pelanggan yang menggunakan *marketplace Blibli* adalah

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

sekitar 10%, dan persentase pelanggan yang menggunakan *marketplace Bukalapak* adalah sekitar 5,7%.

Tabel 2 Jumlah Responden Berdasarkan *Marketplace*

Tipe *Marketplace*	Jumlah Responden	Persentase (%)
Shopee	38	54,4
Tokopedia	12	17,1
Lazada	9	12,8%
Blibli	7	10%
Bukalapak	4	5,7%
Total	**70**	**100**

3.2 Analisis Kuesioner

Pada survei ini, kami mengajukan enam belas pertanyaan kepada 70 orang. Survei yang dilakukan mencakup berbagai topik, seperti kesadaran mereka terhadap pengaruh lingkungan dan sosial, penggunaan *marketplace*, fasilitas untuk melindungi *data* pribadi, dan tujuan mereka menggunakan *marketplace*. Analisis kuesioner yang dilakukan dalam penelitian ini ditunjukkan pada Tabel 3.

3.3 Analisis Hasil Kuesioner

Langkah selanjutnya adalah menguji kuesioner berdasarkan tanggapan responden. Keenam belas pertanyaan kuesioner dievaluasi menggunakan skala Guttman, dengan pilihan jawaban ya/benar dan tidak/salah. Hasilnya diperoleh dengan membagi jumlah responden sebesar 100%, yang didasarkan pada skor ya/tidak kuesioner (Tabel 4). Berikut ini adalah hasilnya:

1. Dari 70 responden yang disurvei, 64,28% menyatakan berbelanja di *marketplace* tergantung keinginannya. Sebaliknya, 35,71% mengaku tidak berbelanja ke *marketplace* karena tidak memenuhi kebutuhannya.
2. Setelah mengumpulkan data dari 70 peserta, kami menemukan bahwa 74,28% dari mereka menggunakan *marketplace*, jumlah ini cukup tinggi mengingat betapa lazimnya *marketplace* dalam lingkungan sosial. Sementara itu, 11,42% mengatakan mereka menggunakan *marketplace* untuk alasan yang tidak berhubungan dengan jangkauan sosialnya.
3. Hasilnya menunjukkan bahwa dari 70 orang yang mengikuti survei, 45,71% menyadari risiko jika tidak melakukan tindakan pencegahan yang memadai untuk mengamankan informasi pribadi mereka saat membuat akun di *marketplace*. Meskipun hal ini terjadi, 54,28% responden mengaku tidak tahu tentang bahaya pelanggaran data.
4. Responden berjumlah 70 orang, dan hasilnya 37,14% berpendapat *marketplace* aman untuk data pribadi dan tidak rentan terhadap perubahan yang dilakukan pihak lain. Di sisi lain, 62% responden mengatakan mereka tidak aman terkait *data* pribadi dan pihak ketiga dapat memodifikasinya.
5. Berdasarkan data yang diperoleh dari 70 responden, 78,57% meyakini bahwa *marketplace* yang mereka gunakan dapat memberikan jaminan keamanan informasi pribadi konsumen. Sementara itu, 21,42% orang yang mengikuti survei mengatakan bahwa *marketplace* yang mereka gunakan tidak memberikan keamanan yang memadai terhadap informasi pribadi mereka.
6. Berdasarkan data yang dihimpun dari 70 responden, 68,57% di antaranya menyatakan *marketplace* tersebut aman dan mudah digunakan. *Marketplace* juga tidak aman dan sulit digunakan, menurut 31,42% responden.
7. Dari data yang diperoleh dari 80 responden, dapat disimpulkan bahwa 17,14% merasa yakin dengan kemampuannya menangani masalah keamanan data pribadi secara mandiri. Sementara itu, 82% responden mengatakan mereka tidak akan mampu mengatasi masalah keamanan data pribadi sendirian.
8. Hasil penelitian menunjukkan bahwa 47,14% orang yang mengikuti survei yakin bahwa *marketplace* mereka melakukan tindakan pencegahan untuk mencegah kebocoran data.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Statistik ini menunjukkan peningkatan, meskipun 47,14% responden mengatakan *marketplace* mereka tidak cukup aman untuk mencegah pelanggaran data.

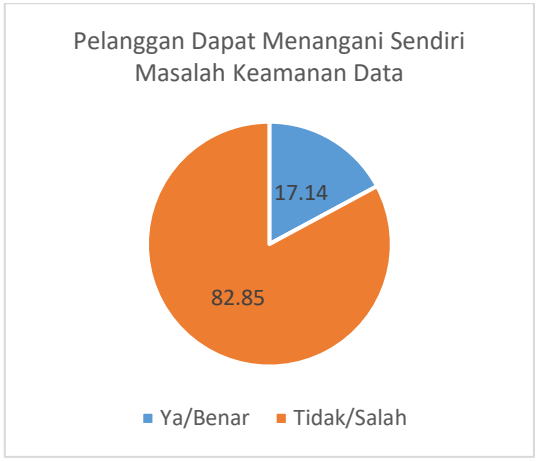
9. Dari 70 orang yang mengikuti survei, 88,57% mengatakan mereka dapat mempercayai *marketplace* untuk menjaga keamanan informasi pribadi mereka. Persentase masyarakat yang merasa informasi pribadinya tidak terlindungi secara memadai oleh *Marketplace* yang mereka gunakan adalah 11,42%.
10. Informasi keamanan data yang diberikan *website marketplace* adalah akurat menurut 90% responden. Persentase responden yang merasa *website marketplace* memberikan informasi keamanan data palsu adalah 10%.
11. Terhitung 88,57% responden (62 dari 70) menyatakan bahwa *marketplace* yang diakses kompatibel dengan perangkatnya, berdasarkan data yang diperoleh dari 70 responden. Selain itu, 11,42% peserta survei mengatakan gadgetnya tidak kompatibel dengan *marketplace* yang digunakan.
12. Data dari 70 responden menunjukkan bahwa 40% dari mereka yang disurvei rutin menggunakan teknologi keamanan *data* di *marketplace* masing-masing. Sistem keamanan data di industri yang mereka gunakan belum dimanfaatkan secara umum oleh 60% responden.
13. Di antara 70 orang yang mengikuti survei, 94,28% menyatakan sering menggunakan *marketplace*. Selain itu, 5,71% masyarakat yang mengikuti survei menyatakan jarang berbelanja di *marketplace*.
14. Berdasarkan data yang dikumpulkan dari 70 peserta, 62% menyatakan selalu menggunakan *marketplace* untuk berbagai hal. Di sisi lain, 37,14% menyatakan tidak selalu memanfaatkan *marketplace* untuk mencapai beberapa tujuan.
15. Data yang dikumpulkan dari 70 responden menunjukkan bahwa 81,42% pengguna bebas menggunakan situs *marketplace* dan memberikan informasi pribadi untuk pengisian akun. Meskipun 18,57% orang yang mengikuti survei mengatakan bahwa mereka menyelesaikan akun mereka menggunakan situs web dan memberikan informasi pribadi untuk alasan selain pilihan mereka sendiri, 13 dari 70 orang mengatakan demikian.
16. Data dari 70 responden menunjukkan bahwa 74,28% responden menyatakan bebas memilih *marketplace* barang bekas setelah melihat-lihat *marketplace* lainnya. Meskipun 25,71% responden mengatakan mereka tidak dapat memilih atau bahkan mempertimbangkan untuk menggunakan *marketplace* lain, namun hal tersebut masih terjadi.

3.4 Visualisasi Data

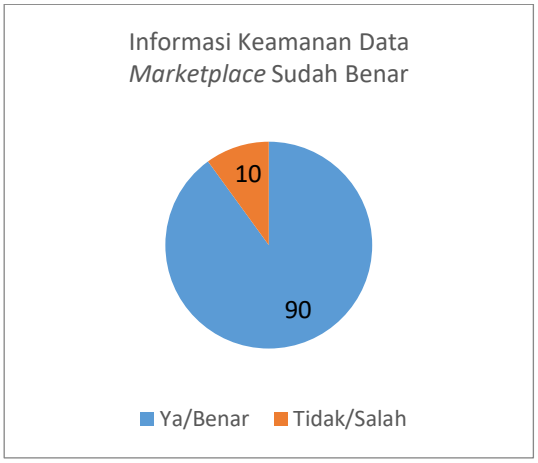
Dalam mendapatkan pemahaman tentang cara responden berperilaku sehubungan dengan perlindungan informasi pribadi pelanggan ketika mereka melakukan pembelian secara online, pendekatan visualisasi data melibatkan pemanfaatan analisis kuesioner untuk memberikan jawaban atas pertanyaan yang memiliki tingkat respons lebih dari 80 persen. Bedasarkan Gambar 3, kita dapat melihat bahwa 82% pelanggan *marketplace* mengatakan mereka tidak dapat memperbaiki masalah keamanan *data* pribadi saat menggunakan aplikasi *marketplace*, sementara 17,14% mengatakan mereka dapat memperbaiki masalah seperti verifikasi *data* pribadi. Informasi keamanan *data* yang diberikan pada saat pendaftaran aplikasi *marketplace* dinilai sangat lengkap dan jelas oleh 90% pengguna sesuai Gambar 4. Sementara itu, 10% pelanggan menyatakan masih belum jelas bagaimana aplikasi *marketplace* melindungi *data* pengguna.

Meskipun 11,42% pelanggan pasar mengeluh bahwa aplikasi mereka tidak kompatibel dengan perangkat *smartphone*, 88,57 persen konsumen menggunakan *marketplace* karena aplikasi tersebut berfungsi dengan perangkat mereka pada Gambar 5. Selanjutnya, Gambar 6 menunjukkan bahwa 5,41% pelanggan pasar melaporkan menggunakan aplikasi sesekali, sementara 94,28% konsumen sering menggunakan fitur dari *marketplace*. Berdasarkan Gambar 7, 81,42% pelanggan *marketplace* dengan bebas memberikan informasi pribadinya saat mendaftar akun. Pada saat yang sama, 18,57% pengguna mengatakan mereka terpaksa mengisi *data* pribadi mereka untuk mendaftar sepenuhnya di *marketplace*.

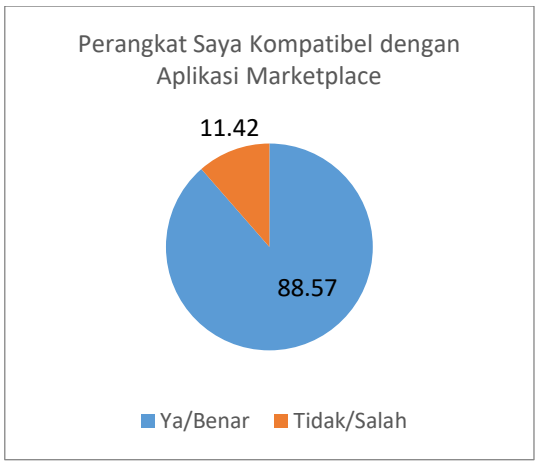
Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.



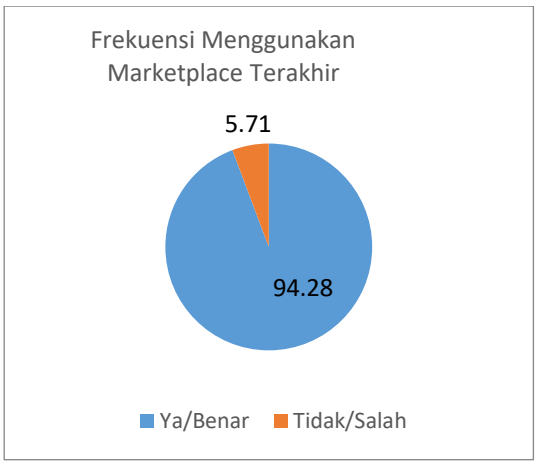
Gambar 3 Pertanyaan No. 7



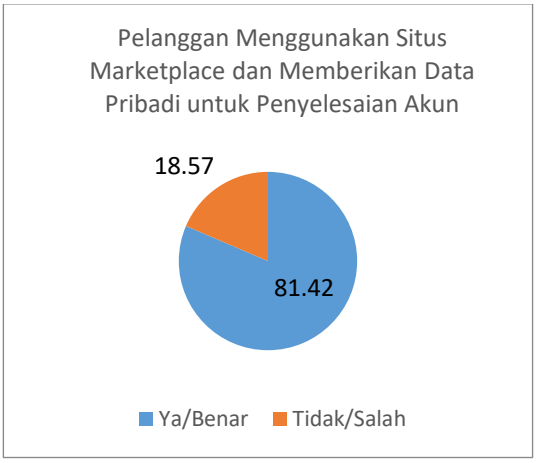
Gambar 4 Pertanyaan No. 10



Gambar 5 Pertanyaan No. 11



Gambar 6 Pertanyaan No. 13



Gambar 7 Pertanyaan No. 15

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Tabel 3 Analisis Kuesioner

No.	Pertanyaan	Aspek
1	Pelanggan menggunakan *marketplace* atas kemauannya sendiri.	Pengaruh Lingkungan Dan Sosial
2	Pelanggan berbelanja di *marketplace* karena populer di komunitas mereka.	
3	Saat mendaftar di *marketplace*, pelanggan menyadari risiko keamanan informasi pribadi mereka.	
4	Ketika data pribadi disimpan di *marketplace*, pelanggan mengklaim data tersebut aman dan tidak diubah oleh pihak luar.	
5	*Marketplace* pilihan pelanggan mungkin menawarkan jaminan keamanan untuk data pribadi pelanggan.	
6	Saat menggunakan *website marketplace* pelanggan, pengguna merasa aman dan nyaman	
7	Pelanggan dapat menyelesaikan sendiri masalah apa pun terkait keamanan informasi pribadinya.	Fasilitas Keamanan Data Pribadi
8	Pelanggan yakin bahwa *marketplace* yang mereka gunakan aman dari kemungkinan pelanggaran data.	
9	Pelanggan mungkin mempercayai perlindungan data pribadi *marketplace*	
10	Informasi situs web *marketplace* mengenai keamanan data akurat.	
11	Perangkat pelanggan dan *marketplace* kompatibel.	
12	Pelanggan puas memanfaatkan teknologi keamanan data yang saat ini tersedia di *marketplace*.	
13	Pelanggan sering berbelanja di *marketplace* pelanggan saat ini.	Tujuan Penggunaan *Marketplace*
14	Pelanggan sering menggunakan *marketplace* untuk berbagai tujuan.	
15	Karena keinginan pribadi, pelanggan mengunjungi situs *marketplace* dan memberikan informasi pribadi untuk penyelesaian akun.	Kesadaran Diri Mengenai Keamanan Data Pelanggan Saat Menggunakan The Aplikasi *Marketplace*
16	Pelanggan memanfaatkan *marketplace* ini sebagai hasil dari pengambilan keputusan secara sadar untuk melakukannya	

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Tabel 4 Hasil Kuesioner

No.	Pertanyaan	Jawaban Responden	
		Ya/Benar	Tidak/Salah
1	Pelanggan menggunakan *marketplace* atas kemauannya sendiri.	45	25
2	Pelanggan berbelanja di *marketplace* karena populer di kalangan mereka.	52	18
3	Saat mendaftar di *marketplace*, pelanggan menyadari risiko keamanan informasi pribadi mereka.	32	38
4	Ketika data pribadi disimpan di *marketplace*, pelanggan mengklaim data tersebut aman dan tidak diubah oleh pihak luar.	26	44
5	*Marketplace* pilihan pelanggan mungkin menawarkan jaminan keamanan untuk data pribadi pelanggan.	55	15
6	Saat menggunakan *website marketplace* pelanggan, pengguna merasa aman dan nyaman.	48	22
7	Pelanggan dapat menyelesaikan sendiri masalah apa pun terkait keamanan informasi pribadinya.	12	58
8	Pelanggan yakin bahwa *marketplace* yang mereka gunakan aman dari kemungkinan pelanggaran data.	33	37
9	Pelanggan mungkin mempercayai perlindungan data pribadi *marketplace*.	62	8
10	Informasi situs web *marketplace* mengenai keamanan data akurat.	63	7
11	Perangkat pelanggan dan *marketplace* kompatibel.	62	8
12	Pelanggan puas memanfaatkan teknologi keamanan data yang saat ini tersedia di *marketplace*.	28	42
13	Pelanggan sering berbelanja di *marketplace* pelanggan saat ini.	66	4
14	Pelanggan sering menggunakan *marketplace* untuk berbagai tujuan.	44	26
15	Karena keinginan pribadi, pelanggan mengunjungi situs *marketplace* dan memberikan informasi pribadi untuk penyelesaian akun.	57	13
16	Pelanggan menggunakan *marketplace* atas kemauannya sendiri.	52	18

4. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa beberapa pelanggan pasar sudah mengetahui dengan baik tentang pentingnya melindungi informasi pribadi mereka ketika mereka mendaftar untuk aplikasi *marketplace*. Pasalnya, *website* atau aplikasi tersebut dengan jelas mencantumkan informasi apa saja yang dibutuhkan pelanggan dalam hal kelengkapan data pribadinya. Mayoritas 81,42% pengguna merasa yakin dengan kemampuan aplikasi *marketplace* yang digunakan dalam menjaga keamanan informasi pribadi mereka. Karena 88,57% konsumen yakin bahwa aplikasi *marketplace* yang mereka gunakan memiliki semacam tindakan pengamanan data, hal ini jelas terkait dengan apa yang dibutuhkan konsumen untuk menggunakan fitur-fitur *marketplace*. orang yang membeli produk. Khususnya, 82% pelanggan mengatakan mereka tidak dapat memperbaiki masalah terkait keamanan aplikasi pasar dan data pribadi mereka. Bagi peneliti, masalah keamanan data pelanggan dalam kaitannya dengan penggunaan *marketplace* merupakan hal yang penting dan relevan, terutama mengingat meluasnya penggunaan data dan teknologi dalam *e-commerce*. Definisi penggunaan keamanan data dan penggunaan data yang benar-benar aman dan terjamin merupakan aspek penting dalam pengembangan dan penggunaan sistem *e-commerce.*

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

DAFTAR PUSTAKA

Adrian, D., & Mulyandi, M. R. (2021). Manfaat Pemasaran Media Sosial Instagram Pada Pembentukan Brand Awareness Toko Online. *Jurnal Indonesia Sosial Sains*, *2*(02), 215–222. https://doi.org/10.59141/JISS.V2I02.195

Ahdiat, A. (2024). *5 E-Commerce dengan Pengunjung Terbanyak Sepanjang 2023*. Katadata. https://databoks.katadata.co.id/datapublish/2024/01/10/5-e-commerce-dengan-pengunjung-terbanyak-sepanjang-2023

Ariska, I., & Amelia, R. (2021). Analisis Tingkat Kepuasan Pengguna Marketplace Shopee dan Lazada Menggunakan Metode End User Computing Satisfaction (EUCS). *Bina Darma Conference on Computer Science (BDCCS)*, *3*(2), 321–327. https://conference.binadarma.ac.id/index.php/BDCCS/article/view/2136

Bakti, U., & Septijantini Alie, M. (2020). Pengaruh Kualitas Pelayanan, Produk dan Harga Terhadap Minat Beli Pada Toko Online Lazada di Bandar Lampung. *JURNAL EKONOMI*, *22*(1), 101–118. https://doi.org/10.37721/JE.V22I1.633

Hasanah, N. (2019). *Analisis Mekanisme Dropshipper dan Reseller di Toko Online S3 Komputer Surabaya* [Universitas Islam Negeri Sunan Ampel Surabaya]. http://digilib.uinsa.ac.id/35394/

Hoofnagle, C. J., van der Sloot, B., & Borgesius, F. Z. (2019). The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, *28*(1), 65–98. https://doi.org/10.1080/13600834.2019.1573501

Iswandari, B. A. (2021). Jaminan Atas Pemenuhan Hak Keamanan Data Pribadi Dalam Penyelenggaraan E-Government Guna Mewujudkan Good Governance. *Jurnal Hukum Ius Quia Iustum*, *28*(1), 115–138. https://doi.org/10.20885/iustum.vol28.iss1.art6

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), 305–311. https://doi.org/10.1038/s42256-020-0186-1

Karina, M. (2019). Pengaruh E-servicescape Online Marketplace Shopee pada Perceived Value dan Kepuasan Pelanggan, serta Dampaknya terhadap Loyalitas Pelanggan. *Jurnal Maksipreneur: Manajemen, Koperasi, Dan Entrepreneurship*, *9*(1), 103. https://doi.org/10.30588/jmp.v9i1.534

Kesuma, A. A. N. D. H., Budiartha, I. N. P., & Wesna, P. A. S. (2021). Perlindungan Hukum Terhadap Keamanan Data Pribadi Konsumen Teknologi Finansial dalam Transaksi Elektronik. *Jurnal Preferensi Hukum*, *2*(2), 411–416. https://doi.org/10.22225/jph.2.2.3350.411-416

Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X., & He, B. (2023). A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), 3347–3366. https://doi.org/10.1109/TKDE.2021.3124599

Lu, Y., Huang, X., Dai, Y., Maharjan, S., & Zhang, Y. (2020). Blockchain and Federated Learning for Privacy-Preserved Data Sharing in Industrial IoT. *IEEE Transactions on Industrial Informatics*, *16*(6), 4177–4186. https://doi.org/10.1109/TII.2019.2942190

Mulyati, Y., & Gesitera, G. (2020). Pengaruh Online Customer Review terhadap Purchase Intention dengan Trust sebagai Intervening pada Toko Online Bukalapak di Kota Padang. *Jurnal Maksipreneur: Manajemen, Koperasi, Dan Entrepreneurship*, *9*(2), 173. https://doi.org/10.30588/jmp.v9i2.538

Prajoko, M. A., Effendi, I., & Sugandini, D. (2022). Pengaruh Persepsi Kegunaan, Kualitas Informasi, Terhadap E-Kepuasan dengan Kepercayaan Sebagai Variabel Mediasi pada Pengguna Marketplace Tokopedia di Daerah Istimewa Yogyakarta. *JMBI UNSRAT (Jurnal Ilmiah Manajemen Bisnis Dan Inovasi Universitas Sam Ratulangi).*, *9*(1). https://doi.org/10.35794/JMBI.V9I1.38987

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, *25*(1), 37–43. https://doi.org/10.1038/s41591-018-0272-7

Prilano, K., Sudarso, A., & Fajrillah, F. (2020). Pengaruh Harga, Keamanan dan Promosi Terhadap Keputusan Pembelian Toko Online Lazada. *Journal of Business and Economics Research (JBE)*, *1*(1), 1–10. https://doi.org/10.47065/jbe.v1i1.56

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Putra, H. F., Wirawan, W., & Penangsang, O. (2019). Penerapan Blockchain dan Kriptografi untuk Keamanan Data pada Jaringan Smart Grid. *Jurnal Teknik ITS*, *8*(1), A11–A16. https://doi.org/10.12962/j23373539.v8i1.38525

Putri, N. I., Komalasari, R., & Munawar, Z. (2020). Pentingnya Keamanan Data dalam Intelijen Bisnis. *J-SIKA | Jurnal Sistem Informasi Karya Anak Bangsa*, *2*(02), 41–48. https://ejournal.unibba.ac.id/index.php/j-sika/article/view/381

Rachmadewi, I. P., Firdaus, A., Qurtubi, Q., Sutrisno, W., & Basumerda, C. (2021). Analisis Strategi Digital Marketing pada Toko Online Usaha Kecil Menengah. *Jurnal INTECH Teknik Industri Universitas Serang Raya*, *7*(2), 121–128. https://doi.org/10.30656/intech.v7i2.3968

Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online JD.ID Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *10*(2), 681–686. https://doi.org/10.24176/SIMET.V10I2.3487

Singarimbun, M., & Effendi, S. (1995). *Metode Penelitian Survei* (2nd ed.). LPPPES. https://digilib-himbaetam.id/index.php?p=show_detail&id=70&keywords=

Suhatman, S., Sari, M. R., Nagara, P., & Nasfi, N. (2020). Pengaruh Atribut Produk dan Promosi Terhadap Minat Beli Konsumen Kota Pariaman di Toko Online Shopee. *Jurnal Bisnis, Manajemen, Dan Ekonomi*, *1*(2), 26–41. https://doi.org/10.47747/jbme.v1i2.81

Ula, M. (2019). Analisis Metode Pengamanan Data pada Layanan Cloud Computing. *TECHSI - Jurnal Teknik Informatika*, *11*(1), 116. https://doi.org/10.29103/techsi.v11i1.1357

Yang, M. Z., & Sihotang, J. I. (2023). Analisis Kepuasan Pengguna Terhadap User Interface Aplikasi E-Commerce Shopee Menggunakan Metode EUCS di Jakarta Barat. *Informatics and Digital Expert (INDEX)*, *4*(2), 53–60. https://doi.org/10.36423/index.v4i2.1110

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024, Pp. 105 – 111
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Deep Learning dalam Prediksi Kebiasaan Merokok di Inggris Guna Mendukung Kebijakan Kesehatan Masyarakat yang Lebih Efektif

Muhammad Arden Prabaswara [(1)], Kalistus Haris Pratama [(2)], Desva Fitranda Majid [(3)], Febri Liantoni [(4)*]
Teknik Informatika dan Komputer, Fakultas Keguruan dan Ilmu Pendidikan,
Universitas Sebelas Maret, Surakarta
e-mail : {ardenio88,hariswonogiri,desvafitranda}@student.uns.ac.id, febri.liantoni@gmail.com.
* Penulis korespondensi.
Artikel ini diajukan 20 Desember 2023, direvisi 20 April 2024, diterima 2 Mei 2024, dan dipublikasikan 25 Mei 2024.

Abstract

Smoking is a common practice throughout the world, where a person smokes and inhales the smoke produced from burning tobacco or other tobacco products. This action has become a significant global health issue because of the various health risks. This activity is often considered an addictive habit because nicotine, the psychoactive compound in tobacco, can cause physical and psychological dependence. This research applies Deep Learning methods to predict data on smoking habits in the UK. The dataset used in this research includes information about gender, age, marital status, highest level of education, nationality, ethnicity, income, and region. Through this research using Deep Learning methods, we can examine a complex data set that describes Smoking Habits in the UK. Based on trials with a dataset of 1,691 items, an accuracy of 78% was obtained. This research can provide important insights into the effectiveness of anti-smoking policies that have been implemented and help plan further actions to reduce the prevalence of smoking and its negative impact on society.

Keywords: Smoke, Predictions, Deep Learning, Tobacco, Addictive

Abstrak

Merokok adalah praktik yang umum di seluruh dunia, di mana seseorang menghisap dan menghirup asap yang dihasilkan dari pembakaran tembakau atau produk tembakau lainnya. Tindakan ini telah menjadi isu kesehatan global yang penting karena berbagai risiko yang mengganggu kesehatan. Aktivitas ini seringkali dianggap sebagai kebiasaan adiktif karena nikotin, senyawa psikoaktif dalam tembakau dapat menyebabkan ketergantungan fisik dan psikologis. Penelitian ini menerapkan metode *deep learning* dalam memprediksi data kebiasaan merokok di Inggris. *Dataset* yang digunakan dalam penelitian ini mencakup informasi tentang jenis kelamin, umur, status pernikahan, pendidikan terakhir, kewarganegaraan, etnis, pendapatan, dan wilayah. Melalui penelitian ini dengan metode *deep learning*, dapat memeriksa kumpulan data yang kompleks yang menggambarkan Kebiasaan Merokok di Inggris. Berdasarkan uji coba dengan *dataset* 1.691 *item*, diperoleh akurasi sebesar 78%. Penelitian ini dapat memberikan wawasan penting tentang efektivitas kebijakan anti-merokok yang telah diterapkan dan membantu merencanakan tindakan selanjutnya untuk mengurangi prevalensi merokok dan dampak negatifnya terhadap masyarakat.

Kata Kunci: Merokok, Prediksi, *Deep Learning*, Tembakau, Adiktif

1. PENDAHULUAN

Di Britania Raya (UK), pemerintah telah mengimplementasikan peraturan publik yang ketat guna menjaga kesehatan masyarakat terkait konsumsi rokok. Salah satu langkah kunci adalah larangan merokok di dalam ruang publik yang tertutup, termasuk restoran, sarana transportasi umum, dan tempat kerja (Akinosho et al., 2020; Jones et al., 2015). Selain itu, kemasan rokok wajib menampilkan peringatan kesehatan yang mencolok, dan iklan produk tembakau dibatasi secara ketat. Upaya lainnya yang dilakukan oleh UK adalah dengan meningkatkan tarif cukai tembakau, yang bertujuan untuk mengurangi konsumsi rokok serta mendorong perokok untuk berhenti, dimana itu semua merupakan bagian dari inisiatif pemerintah guna mengurangi dampak

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

negatif rokok terhadap kesehatan masyarakat (Allender et al., 2009; Najafabadi et al., 2015). Oleh karena itu, penting untuk memahami faktor-faktor yang mempengaruhi kebiasaan merokok, mengembangkan cara untuk memprediksi, dan mengurangi tren merokok untuk meningkatkan kesehatan masyarakat.

Di Inggris, seperti di banyak negara lainnya, berbagai upaya telah dilakukan untuk mengurangi jumlah perokok. Namun untuk memperoleh tujuan tersebut, diperlukan pemahaman yang lebih mendalam mengenai faktor-faktor yang menyebabkan individu mulai atau terus merokok. Secara internasional, pemerintah berinvestasi dalam upaya untuk mencoba dan meningkatkan pengambilan keputusan program dan kebijakan kesehatan masyarakat (Rickert et al., 2007). Penggunaan bukti penelitian dalam pengambilan keputusan kebijakan kesehatan masyarakat dipengaruhi oleh serangkaian faktor kontekstual yang terjadi pada tingkat individu, organisasi, dan eksternal. Lalu untuk prediksi kebiasaan merokok sendiri menjadi penting dalam konteks kebijakan kesehatan karena pemahaman terhadap kebiasaan merokok berperan penting dalam merencanakan strategi pencegahan yang efektif untuk mengurangi dampak penyakit dan kematian akibat merokok.

Prediksi kebiasaan merokok membantu alokasi sumber daya kesehatan yang lebih efisien, sementara upaya pencegahan dan dukungan bagi individu untuk berhenti merokok dapat meningkatkan kualitas hidup penduduk. Faktanya, status merokok sebelumnya dan pengaruh teman memiliki peran penting sebagai prediktor kebiasaan merokok. Keberlanjutan perilaku merokok dari masa remaja hingga dewasa menekankan pentingnya program pencegahan di sekolah menengah pertama, namun juga perlu diperhatikan bahwa banyak individu mulai merokok setelah itu. Oleh karena itu, pemahaman dan prediksi perilaku merokok mendukung desain kebijakan kesehatan yang efektif, melalui identifikasi faktor risiko yang terkait dengan merokok, pengembangan program pencegahan yang ditargetkan pada kelompok berisiko tinggi, dan alokasi sumber daya kesehatan yang lebih efisien (Rickert et al., 2007).

Beberapa tahun terakhir, kemajuan teknologi dan perkembangan kecerdasan buatan khususnya metode *deep learning* yang telah membuka peluang baru dalam menganalisis dan memprediksi perilaku manusia. *Deep learning* adalah aspek jaringan saraf tiruan yang bertujuan untuk meniru teknik *machine learning* yang digunakan manusia untuk memperoleh jenis pengetahuan tertentu. *Deep learning* dapat didefinisikan sebagai mempelajari berbagai tingkat representasi dan abstraksi yang membantu kita memahami data seperti gambar, suara, dan teks (Karlsson et al., 2021; Ristoski et al., 2015; Sathishkumar et al., 2023). *Deep learning* memiliki kemampuannya untuk secara otomatis menggali pola-pola kompleks dari data melalui jaringan *neural multi-layer*, memungkinkan analisis efisien terhadap data berskala besar, generalisasi yang kuat terhadap situasi baru, dan kemampuan untuk menangani data dalam format yang beragam seperti gambar, suara, serta teks. Perbedaan mendasar dibandingan *machine learning* biasa terletak pada kemampuannya untuk mengekstraksi fitur-fitur yang abstrak, algoritma optimasi yang lebih canggih, dan ketergantungan pada sumber daya komputasi yang besar untuk pelatihan model yang optimal. Metode ini dapat digunakan sebagai alat yang sangat efektif dalam penelitian untuk memprediksi kebiasaan merokok. *Deep learning* mampu memproses data kompleks dan mengekstraksi fitur yang mendalam, yang memungkinkan peneliti untuk mengembangkan model prediktif yang lebih kuat dan akurat dalam memahami faktor-faktor yang mempengaruhi kebiasaan merokok.

Artikel ini menggunakan penggunaan *deep learning* sebagai salah satu alat untuk memprediksi kebiasaan merokok di Inggris dengan *dataset* yang digunakan dalam penelitian ini mencakup informasi tentang jenis kelamin, umur, status pernikahan, pendidikan terakhir, kewarganegaraan, etnis pendapatan, dan wilayah. Dengan memanfaatkan *big data* dan teknik pemrosesan data tingkat lanjut, kita dapat mengidentifikasi pola mendasar perilaku merokok dan faktor-faktor yang mempengaruhi kebiasaan tersebut. Selain itu, pentingnya mengevaluasi efektivitas kebijakan anti-merokok yang telah diterapkan di Inggris merupakan faktor yang signifikan dalam upaya mengurangi jumlah perokok. Proses evaluasi ini dapat memberikan informasi berharga mengenai keberhasilan upaya pencegahan merokok yang dilaksanakan atau upaya yang perlu ditingkatkan.

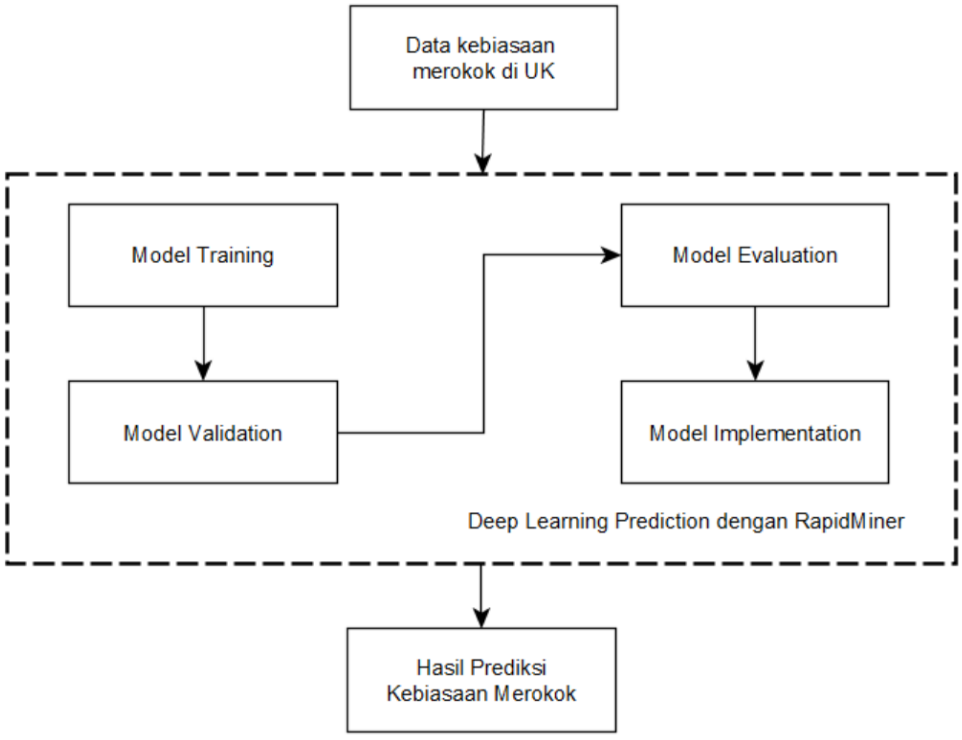
Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Oleh karena itu, penelitian ini bertujuan untuk menjelaskan konsep dasar *deep learning*, menyelidiki peran teknologi ini dalam mempelajari kebiasaan merokok, dan mengeksplorasi potensi penggunaannya dalam memprediksi tren merokok di Inggris. Dengan menggunakan pendekatan *deep learning*, diharapkan hasil penelitian ini dapat memberikan gambaran yang lebih komprehensif mengenai kebiasaan merokok di Inggris dan memungkinkan pengambilan keputusan yang lebih efektif yang bertujuan untuk mengurangi angka perokok di Inggris.

2. METODE PENELITIAN

Penelitian ini dilakukan dengan tujuan untuk memprediksi kebiasaan merokok pada individu tertentu di Inggris. Metode ini dijadikan sebagai alat untuk menilai apakah seseorang cenderung merokok atau tidak, sehingga dapat memberikan kontribusi berharga dalam merancang kebijakan kesehatan masyarakat, terutama dalam upaya anti-merokok. Dengan memahami faktor-faktor yang mempengaruhi kebiasaan merokok, penelitian ini diharapkan dapat membantu masyarakat untuk mengurangi prevalensi merokok dan dampak negatifnya terhadap kesehatan.

Penelitian ini menggunakan metode *deep learning* untuk analisis data. Peneliti menerapkan jaringan saraf tiruan yang mendalam (*deep neural networks*) dengan bantuan perangkat lunak RapidMiner. RapidMiner merupakan sebuah platform modern untuk analisis yang memiliki beragam fitur seperti *data mining*, analisis prediktif, bisnis *analytics*, *machine learning*, *text mining*, dan banyak lagi (Čižiks & Grabusts, 2019; Kotu & Deshpande, 2019; LeCun et al., 2015). Dalam konteks ini, RapidMiner digunakan untuk mengukur dan mengoptimalkan kinerja algoritma-algoritma *deep learning*. Tujuannya adalah menemukan algoritma terbaik untuk keperluan klasifikasi, prediksi, dan teknik *data mining* lainnya yang tidak hanya efektif tapi juga mudah digunakan (Čižiks & Grabusts, 2019; Zardo & Collie, 2014). Di samping itu, RapidMiner juga dilengkapi dengan fitur-fitur lain seperti analisis dan visualisasi prediktif, evaluasi data, manipulasi data, pembuatan model, dan sebagainya (Mccaig, 1990; Zardo & Collie, 2014). Langkah-langkah penelitian dijelaskan dalam diagram alur seperti yang terlihat pada Gambar 1.



Gambar 1 Alur Metode *Deep Learning* dengan RapidMiner

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Pada tahap *preprocessing* data, terdapat langkah penting untuk memastikan validitas dan kualitas *dataset* yang terdiri dari 1691 *item*. Pembersihan data, juga dikenal sebagai *data cleaning*, merupakan langkah awal yang penting. Langkah-langkah untuk melakukan proses ini, tujuannya adalah untuk dapat memverifikasi keakuratan serta relevansi data yang diterapkan dalam analisis, perlu dipastikan bahwa data tersebut sesuai dengan keperluan riset yang bertujuan untuk memprediksi perilaku merokok. Pada tahap ini, data yang tidak akurat atau tidak lengkap diidentifikasi dan dihilangkan dengan menggunakan berbagai metode yang sesuai dengan jenis data yang diproses. Jika kumpulan data ini berisi data numerik, peneliti memeriksa rentang nilai yang valid, mengidentifikasi dan menangani nilai yang hilang, serta mendeteksi dan menangani *outlier* sesuai dengan metode peneliti dengan manajemen data yang baik. Langkah ini penting karena akan memastikan bahwa *dataset* yang akan digunakan untuk melatih model *deep learning* bersih, andal, dan siap untuk dianalisis secara mendalam guna mendapatkan hasil prediksi yang akurat.

Proses prediksi kebiasaan merokok menggunakan metode *deep learning*, yang telah terbukti memiliki akurasi dan kapabilitas yang tinggi ketika diterapkan pada *dataset* berukuran besar, seperti yang dijelaskan dalam penelitian sebelumnya (Kitcharoen et al., 2013). Metode ini berlandaskan pada jaringan saraf tiruan (*artificial neural network*) yang dirancang untuk mempelajari representasi data yang kompleks dan mengekstraksi fitur mendalam yang memungkinkan model untuk memahami faktor-faktor yang mempengaruhi kebiasaan merokok. Melalui pendekatan ini, kita dapat memprediksi kebiasaan merokok dengan menggabungkan informasi awal atau data latihan dengan bukti baru atau data uji dalam suatu model yang kuat berdasarkan *deep learning*.

3. HASIL DAN PEMBAHASAN

Kumpulan data yang diperoleh pada langkah pemilihan data dipisahkan dengan metode validasi silang. Validasi silang merupakan pendekatan yang kuat untuk menguji performa model. Dalam konteks ini, data akan dibagi menjadi lipatan (*folds*) yang lebih kecil, dan proses validasi akan dilakukan pada setiap lipatan. Khususnya, peneliti akan menggunakan validasi silang 10-*folds*, di mana data akan dibagi menjadi 10 lipatan. Pada setiap iterasi, satu lipatan akan diambil sebagai data uji, sementara sembilan lipatan lainnya akan digunakan sebagai data latih. Dengan demikian, model akan diuji sebanyak 10 kali dan hasil performanya akan diambil sebagai indikator yang kuat untuk akurasi model dalam prediksi kebiasaan merokok.

Tabel 1 menunjukkan contoh data awal sebelum proses pra-pemrosesan data. Fitur yang digunakan dalam *dataset* ini mencakup beberapa aspek terkait prediksi kebiasaan merokok. Dalam proses analisis data ini, kita akan memastikan bahwa *dataset* yang digunakan dalam pelatihan model *deep learning* telah mengalami tahap pra-pemrosesan dengan baik sehingga dapat menghasilkan hasil prediksi yang lebih akurat dan relevan dengan kebijakan anti-merokok yang sedang diinvestigasi. Fitur-fitur tersebut bisa mencakup *gender*, *age*, *marital status*, *highest qualification*, *nationality*, *ethnicity*, *gross income*, dan *region*.

Tabel 1 Contoh Data Sebelum *Preprocessing*

gender	age	marital_status	highest_qualification	nationality	ethnicity	gross_income	region	smoke
Male	65	Married	Degree	British	White	28,600 to 36,400	Midlands & East Anglia	?
Female	39	Single	Degree	Irish	White	20,800 to 28,600	Midlands & East Anglia	?
Female	57	Married	Higher/Sub Degree	British	White	15,600 to 20,800	Midlands & East Anglia	?
Female	31	Married	A Levels	British	White	2,600 to 5,200	Midlands & East Anglia	?
Male	79	Married	No Qualification	British	White	10,400 to 15,600	Midlands & East Anglia	?

Setelah mengumpulkan data awal, langkah pertama yang dilakukan adalah melakukan praproses data. Hal ini dilakukan untuk mempersiapkan data sebelum melanjutkan analisis lebih lanjut terhadap perkiraan binomial. Sebelum data dapat diolah lebih lanjut, perlu dilakukan konversi ke

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

format yang memenuhi persyaratan analitis, dan akan digunakan model peramalan binomial. Transformasi data ini dilakukan berdasarkan tipe kelas atau nilai kelas pada kumpulan data. Tujuan utama transformasi ini adalah untuk memastikan bahwa data dapat diinterpretasikan dengan benar dan memenuhi persyaratan yang disyaratkan oleh model peramalan binomial. Dengan melakukan langkah ini diharapkan data dapat dipersiapkan secara optimal sebelum dilanjutkan ke analisis lebih lanjut yang melibatkan peramalan binomial.

Proses transformasi data berperan penting dalam meningkatkan kualitas data dan memastikan bahwa persyaratan yang disyaratkan oleh model peramalan binomial terpenuhi. Dengan melakukan transformasi data, peneliti dapat memastikan bahwa model prediksi biner dapat beroperasi dengan efisiensi dan akurasi tinggi saat membuat prediksi pilihan (ya/tidak). Contoh hasil transformasi data dapat dilihat pada Tabel 2 yang menunjukkan bagaimana data ditransformasikan sesuai kebutuhan analitis untuk mencapai prediksi binomial.

Tabel 2 Contoh Data Setelah *Preprocessing*

gender	age	marital_status	highest_qualification	nationality	ethnicity	gross_income	region	smoke
Male	65	Married	Degree	British	White	28,600 to 36,400	Midlands & East Anglia	No
Female	39	Single	Degree	Irish	White	20,800 to 28,600	Midlands & East Anglia	No
Female	57	Married	Higher/Sub Degree	British	White	15,600 to 20,800	Midlands & East Anglia	No
Female	31	Married	A Levels	British	White	2,600 to 5,200	Midlands & East Anglia	No
Male	79	Married	No Qualification	British	White	10,400 to 15,600	Midlands & East Anglia	No

Langkah setelah melakukan persiapan data adalah mengolah data menggunakan alat analisis seperti RapidMiner 10.2 untuk membuat model *forecasting* binomial. RapidMiner adalah perangkat lunak analisis data yang kuat dan terkenal yang menyediakan berbagai algoritma *machine learning* serta teknik analisis untuk pemodelan dan analisis data.

Setelah membuat model *forecasting* binomial menggunakan RapidMiner, langkah selanjutnya adalah menjalankan eksperimen untuk mengevaluasi performa model. Pengujian menggunakan 24 sampel data latih dan data uji yang terpisah. Data eksperimen ini memvalidasi kemampuan model *forecasting* binomial untuk memprediksi hasil secara akurat dan konsisten berdasarkan data dunia nyata. Contoh data eksperimen ditunjukkan pada Tabel 3.

Tabel 3 Contoh Data Uji

gender	age	marital_status	highest_qualification	nationality	ethnicity	gross_income	region	smoke
Male	56	Married	No Qualification	English	White	10,400 to 15,600	Midlands & East Anglia	No
Male	34	Single	GCSE/O Level	British	White	15,600 to 20,800	Midlands & East Anglia	Yes
Female	24	Married	A Levels	English	White	20,800 to 28,600	Midlands & East Anglia	No
Female	42	Single	GCSE/O Level	British	White	15,600 to 20,800	Midlands & East Anglia	No
Male	71	Married	No Qualification	English	White	5,200 to 10,400	Midlands & East Anglia	No

Dari hasil pengujian terhadap 24 sampel data uji, diperoleh akurasi model *forecasting* binomial menggunakan *deep learning* sekitar 78%. Dibandingkan dengan penggunaan *deep learning* pada penelitian lain yang memiliki tingkat akurasi 88%, penelitian peneliti masih memerlukan improvisasi pada pengambilan data dan kualitas data untuk kedepannya, selain kualitas data, pengolahan data juga diperlukan untuk memperbaiki tingkat akurasi (Tweed et al., 2012). Meskipun tingkat akurasi ini mungkin tidak mencapai tingkat yang tinggi, namun tetap relevan dan memiliki dampak positif pada tujuan kita untuk mendukung kampanye anti-merokok di Inggris dan kebijakan kesehatan masyarakat.

Dengan menggunakan model prediksi yang memiliki tingkat akurasi yang relevan, kita dapat membuat keputusan yang lebih baik dalam konteks *forecasting* binomial untuk kebiasaan

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

merokok. Hal ini membantu kita menghindari keputusan yang tidak sesuai dengan realitas, serta memberikan landasan yang lebih kuat untuk mengambil langkah-langkah yang mendukung upaya mengurangi prevalensi merokok dan dampak negatifnya terhadap masyarakat.

4. KESIMPULAN

Berdasarkan hasil serangkaian eksperimen dan evaluasi yang dikembangan dengan metode *forecasting* binomial yang inovatif dengan menggunakan *deep learning* untuk memprediksi kebiasaan merokok di Inggris, diperoleh akurasi mencapai sekitar 78%. Hasil ini menunjukkan bahwa metode yang diterapkan dapat menghasilkan prediksi yang relevan dalam konteks *forecasting* binomial, yang dapat digunakan untuk mendukung kebijakan kesehatan masyarakat yang lebih efektif. Dengan tingkat ketelitian yang sesuai, informasi ini dapat membantu merancang strategi pencegahan yang lebih efektif, mengidentifikasi faktor-faktor risiko yang terkait dengan merokok, serta mengalokasikan sumber daya kesehatan dengan lebih efisien. Dengan hasil penelitian ini diharapkan dapat mendukung kampanye anti-merokok di Inggris dan kebijakan kesehatan masyarakat yang bertujuan mengurangi prevalensi merokok dan dampak negatifnya.

DAFTAR PUSTAKA

Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O., & Ahmed, A. A. (2020). Deep learning in the construction industry: A review of present status and future innovations. *Journal of Building Engineering*, *32*, 101827. https://doi.org/10.1016/j.jobe.2020.101827

Allender, S., Balakrishnan, R., Scarborough, P., Webster, P., & Rayner, M. (2009). The burden of smoking-related ill health in the UK. *Tobacco Control*, *18*(4), 262–267. https://doi.org/10.1136/tc.2008.026294

Čižiks, J., & Grabusts, P. (2019). Data Processing Using The Ibm Spss Modeler Tool. *HUMAN. ENVIRONMENT. TECHNOLOGIES. Proceedings of the Students International Scientific and Practical Conference*, *23*, 16. https://doi.org/10.17770/het2019.23.4388

Jones, A. M., Laporte, A., Rice, N., & Zucchelli, E. (2015). Do Public Smoking Bans have an Impact on Active Smoking? Evidence from the UK. *Health Economics*, *24*(2), 175–192. https://doi.org/10.1002/hec.3009

Karlsson, A., Ellonen, A., Irjala, H., Väliaho, V., Mattila, K., Nissi, L., Kytö, E., Kurki, S., Ristamäki, R., Vihinen, P., Laitinen, T., Ålgars, A., Jyrkkiö, S., Minn, H., & Heervä, E. (2021). Impact of deep learning-determined smoking status on mortality of cancer patients: never too late to quit. *ESMO Open*, *6*(3), 100175. https://doi.org/10.1016/j.esmoop.2021.100175

Kitcharoen, N., Kamolsantisuk, S., Angsomboon, R., & Achalakul, T. (2013). RapidMiner framework for manufacturing data analysis on the cloud. *The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 149–154. https://doi.org/10.1109/JCSSE.2013.6567336

Kotu, V., & Deshpande, B. (2019). Getting Started with RapidMiner. In *Data Science* (pp. 491–521). Elsevier. https://doi.org/10.1016/B978-0-12-814761-0.00015-0

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Mccaig, C. D. (1990). Electric Fields in Vertebrate Repair. *Experimental Physiology*, *75*(2), 280–281. https://doi.org/10.1113/expphysiol.1998.sp004170

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. https://doi.org/10.1186/S40537-014-0007-7/METRICS

Rickert, W. S., Wright, W. G., Trivedi, A. H., Momin, R. A., & Lauterbach, J. H. (2007). A comparative study of the mutagenicity of various types of tobacco products. *Regulatory Toxicology and Pharmacology*, *48*(3), 320–330. https://doi.org/10.1016/j.yrtph.2007.05.003

Ristoski, P., Bizer, C., & Paulheim, H. (2015). Mining the Web of Linked Data with RapidMiner. *Journal of Web Semantics*, *35*, 142–151. https://doi.org/10.1016/j.websem.2015.06.004

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Sathishkumar, V. E., Cho, J., Subramanian, M., & Naren, O. S. (2023). Forest fire and smoke detection using deep learning-based learning without forgetting. *Fire Ecology*, *19*(1), 1–17. https://doi.org/10.1186/S42408-022-00165-0/FIGURES/5

Tweed, J. O., Hsia, S. H., Lutfy, K., & Friedman, T. C. (2012). The endocrine effects of nicotine and cigarette smoke. *Trends in Endocrinology & Metabolism*, *23*(7), 334–342. https://doi.org/10.1016/j.tem.2012.03.006

Zardo, P., & Collie, A. (2014). Predicting research use in a public health policy environment: results of a logistic regression analysis. *Implementation Science : IS*, *9*(1), 142. https://doi.org/10.1186/S13012-014-0142-8/TABLES/2

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024, Pp. 112 – 122
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Analisis dan Optimalisasi Performa Algoritma Gaussian Naive Bayes pada Prediksi *Metabolic Syndrome* Menggunakan SMOTE

Nadiyah Jihan Fauziyah [(1)*], Fadilla Rahmania [(2)], Muhammad Daniyal [(3)], Nur Fitriyah Ayu Tunjung Sari [(4)]
Teknik Informatika, Fakultas Sains dan Teknologi, UIN Maulana Malik Ibrahim, Malang
e-mail : {nadiyahjihanf,fdrahmania1,asus.daniyal}@gmail.com, nur.fitriyah@ti.uin-malang.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 28 Januari 2024, direvisi 30 April 2024, diterima 2 Mei 2024, dan dipublikasikan 25 Mei 2024.

Abstract

Metabolic syndrome is a complex global health problem, with symptoms such as abdominal obesity, insulin resistance, high blood pressure, high blood sugar, and abnormal blood lipids. With this global challenge, several studies have attempted to predict these diseases using machine learning methods. However, often, predictions about a disease result in data imbalance where minority classes are underrepresented. To balance the class proportions, the Synthetic Minority Over-sampling Technique (SMOTE) method replicates the minority class samples. In this research, the technique applied to predict is the Gaussian Naive Bayes (GNB) algorithm. The results show an increase in prediction accuracy by 0.2 from 0.81 to 0.83. This study confirms the critical role of the SMOTE oversampling method in machine learning using the Gaussian Naive Bayes (GNB) algorithm in Metabolic Syndrome prediction and its positive impact on diagnostic efficiency and public health.

Keywords: Metabolic Syndrome, Machine Learning,* Gaussian Naive Bayes*, Synthetic Minority Over-sampling Technique (SMOTE), Prediction

Abstrak

Sindrom Metabolik merupakan masalah kesehatan global yang kompleks, dengan gejala seperti obesitas abdominal, resistensi insulin, tekanan darah tinggi, gula darah tinggi, dan lipid darah abnormal. Menghadapi tantangan global ini, beberapa penelitian telah berusaha memprediksi penyakit ini dengan menggunakan metode pembelajaran mesin. Namun, seringkali prediksi tentang sebuah penyakit menghasilkan ketidakseimbangan data di mana kelas minoritas kurang terwakili. Dalam menyeimbangkan proporsi kelas, metode Synthetic Minority Over-sampling Technique (SMOTE) digunakan dengan mereplikasi sampel kelas minoritas. Dalam penelitian ini, metode yang diterapkan untuk memprediksi adalah algoritma Gaussian Naive Bayes (GNB). Hasilnya menunjukkan peningkatan akurasi prediksi sebesar 0,2 dari yang awalnya 0,81 menjadi 0,83. Penelitian ini menegaskan peran penting metode oversampling SMOTE pada pembelajaran mesin menggunakan algoritma Gaussian Naive Bayes (GNB) dalam prediksi Sindrom Metabolik dan serta dampak positifnya terhadap efisiensi diagnostik dan kesehatan masyarakat.

Kata Kunci: Sindrom Metabolik, Pembelajaran Mesin, Gaussian Naive Bayes, Teknik Pengambilan Sampel Minoritas Sintetis (SMOTE), Prediksi

1. PENDAHULUAN

Penyakit tidak menular, terutama Sindrom Metabolik telah menjadi penyebab kematian global yang signifikan (World Health Organization, 2020). Sindrom Metabolik atau *Metabolic Syndrome* (MetS) adalah kondisi medis yang didiagnosis ketika seseorang memiliki sejumlah faktor risiko yang meningkat untuk penyakit jantung, diabetes tipe 2, dan penyakit pembuluh darah lainnya. Sindrom Metabolik dikenal sebagai himpunan gejala yang mencakup obesitas abdominal, resistensi insulin, tekanan darah tinggi, kadar gula darah tinggi, dan kadar lipid darah abnormal (Huang, 2009). Sindrom ini juga dikenal sebagai sindrom X, resistensi insulin, dll. Dalam literatur, sebenarnya ini bukan penyakit tunggal namun merupakan kumpulan faktor risiko penyakit kardiovaskular dan didefinisikan sedikit berbeda oleh berbagai organisasi (Saklayen, 2018).

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Angka kejadian sindrom metabolik seringkali sejajar dengan kejadian obesitas dan kejadian diabetes tipe 2 yang merupakan salah satu akibat dari sindrom metabolik (Palaniappan et al., 2011). Menurut survei obesitas global di 195 negara yang dilakukan pada tahun 2015, 604 juta orang dewasa dan 108 juta anak-anak mengalami obesitas. Sejak tahun 1980, prevalensi obesitas meningkat dua kali lipat di 73 negara dan meningkat di sebagian besar negara lainnya. Kekhawatiran yang lebih besar adalah bahwa tingkat peningkatan obesitas pada masa kanak-kanak bahkan lebih tinggi (The GBD 2015 Obesity Collaborators, 2017). Di Indonesia sendiri, kondisi ini memiliki dampak signifikan pada kesehatan masyarakat, dengan sekitar 21,66% populasi yang didiagnosis menderita Sindrom Metabolik (Herningtyas & Ng, 2019).

Sindrom metabolik ditandai dengan masalah yang berhubungan dengan obesitas, menunjukkan adanya hubungan antara obesitas dan sindrom metabolik (Han & Lean, 2016). Faktor risiko sindrom metabolik antara lain peningkatan lingkar pinggang atau lemak perut, tingginya trigliserida plasma, peningkatan tekanan darah, gula darah tinggi, dan rendahnya *high-density lipoprotein* (HDL) (Rochlani et al., 2017). Jika seorang pasien memiliki tiga dari lima faktor risiko utama, maka pasien tersebut dikatakan mengalamai sindrom metabolik (Dobrowolski et al., 2022). Beberapa penelitian menunjukkan bahwa risiko sindrom metabolik dapat dibalik secara signifikan dengan mengurangi berat badan dan memfokuskan intervensi pada perubahan pola makan seperti pembatasan waktu makan, pola makan khusus seperti pola makan Mediterania, termasuk meningkatkan latihan fisik, mengubah pola tidur, atau bahkan mengurangi stress yang mengakibatkan sindrom metabolik (Wilkinson et al., 2020).

Beberapa pernyataan mengenai Sindrom Metabolik sebelumnya mendorong kebutuhan akan deteksi dini dan pengelolaan yang efektif. Penyakit kardiovaskular, diabetes, dan komplikasi kesehatan lainnya dapat dihindari atau dikelola lebih baik dengan adanya pendekatan preventif (Han & Lean, 2016). Keterbatasan data dan tantangan dalam prediksi Sindrom Metabolik menuntut eksplorasi terhadap algoritma-algoritma baru yang dapat memberikan kontribusi signifikan. Dari beberapa kasus prediksi penyakit, terdapat beberapa metode yang bisa digunakan, di antaranya *Logitic Regression*, *Gradient Boosting Machine* (GBM)*, K-Nearest Neighbors* (KNN), *Decision Trees* dan *Random Forests*, serta *Gaussian Naïve Bayes* (Zhou et al., 2022).

Dalam penelitian Hu et al. (2022) mereka melakukan prediksi model terhadap 2.714 (30,3%) peserta yang didiagnosis menderita sindrom metabolik. Evaluasi kinerja model menggunakan *Light Gradient Boosting Machine* (LGBM) menunjukkan hasil yang mengesankan. Model pertama memiliki nilai area *under the curve* (AUC) sebesar 0,993, sementara model kedua menunjukkan nilai AUC sebesar 0,885. Meskipun demikian, Model 3 memiliki nilai AUC sebesar 0,859, yang mendekati nilai AUC dari model kedua. Selain itu, nilai AUC untuk model *Logistic Regression* (LR) 1 dan 2 dalam skenario di rumah sakit, serta Model 3 di rumah masing-masing, adalah 0,938, 0,839, dan 0,820.

Penelitian Tavares et al. (2022) juga membahas tentang prediksi sindrom metabolik menggunakan model *machine learning* untuk memprediksi seperti*, Logistic Regression*, *Linear Discriminant Analysis*, *K-Nearest Neighbors (KNN)*, *Decision Trees*, *Light Gradient Boosting Machine (LGBM)*, dan *Extreme Gradient Boosting*. Semua model menunjukkan kalibrasi yang memadai dan diskriminasi yang baik, namun LGBM menunjukkan kinerja yang lebih baik (Sensitivitas = 87,8%, Spesifisitas = 70,2%, AUC-ROC = 0,86). Analisis inferensi kausal menunjukkan bahwa peningkatan tingkat aktivitas fisik dan pengurangan BMI setidaknya sebesar 2% memiliki efek pada pengurangan probabilitas prediksi sindrom metabolik sebesar 3,8% (95% CI = -4,8%; -2,7%).

Pada penelitian ini akan mengolah data yang bersumber di kaggle.com dan mengimplementasikan algoritma Gaussian Naive Bayes yang merupakan algoritma pembelajaran mudah yang memanfaatkan aturan *Bayes* dengan premis atau asumsi tinggi yang karakteristiknya bergantung pada kemandirian yang diberikan oleh kelas (Anand et al., 2022).

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Pemilihan algoritma ini didasarkan pada kebutuhan untuk mendapatkan model yang optimal dalam menangani masalah kompleks yang terkait dengan prediksi sindrom metabolik.

Penelitian Venkata & Pandya (2022) menggunakan algoritma Gaussian Naive Bayes dalam kasus prediksi yang membandingkan beberapa model pembelajaran mesin. Dari hasil penelitian yang menerapkan algoritma *Gaussian Naïve Bayes* menghasilkan akurasi sebesar 99,66% dan standar deviasi sebesar 0,86% ketika diuji dengan menggunakan *k-means cross validation*. Sehingga pengujian secara probabilitas juga cocok untuk prediksi sindrom metabolik.

Selain itu juga terdapat penelitian Libnao et al. (2023) membuat sistem prediksi dan klasifikasi insiden lalu lintas, yang memanfaatkan algoritma *Naïve Bayes*, telah mendapat persetujuan dari Otoritas Pembangunan Metropolitan Manila karena tingkat akurasinya yang tinggi. Penelitian tersebut berhasil menunjukkan kemampuannya dalam memprediksi dan mengklasifikasikan insiden lalu lintas di Metro Manila dengan tingkat akurasi yang mengesankan sebesar 70,03%, menggunakan kumpulan data sebanyak 8.891 catatan.

Dalam masalah di atas terdapat akurasi yang lebih rendah daripada penelitian sebelumnya. Hal ini dapat dilakukan *oversampling* pada data yang digunakan pada *machine learning*. Kombinasi metode yang melakukan *over-sampling* pada kelas minoritas (abnormal) dan *under-sampling* pada kelas mayoritas (normal) dapat mencapai kinerja pengklasifikasi yang lebih baik (dalam ruang ROC) daripada hanya melakukan *under-sampling* pada kelas mayoritas. Selain itu, metode *over-sampling* kelas minoritas dan *under-sampling* kelas mayoritas dapat mencapai kinerja pengklasifikasi yang lebih baik (dalam ruang ROC) daripada memvariasikan rasio kerugian di Ripper atau kelas *prior* di Naive Bayes (Chawla et al., 2002).

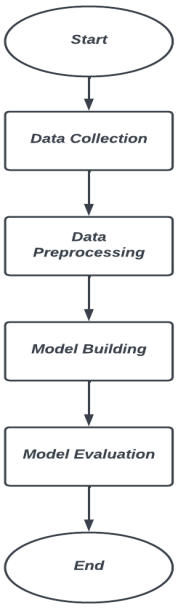
Penelitian sebelumnya menunjukkan bahwa beberapa teknik *machine learning*, termasuk algoritma Gaussian Naive Bayes, memiliki potensi untuk memprediksi sindrom metabolik dengan tingkat akurasi yang memuaskan. Namun, penelitian sebelumnya belum sepenuhnya mengeksplorasi metode yang efektif untuk mengatasi ketidakseimbangan kelas dalam data sindrom metabolik dan bagaimana hal ini dapat memengaruhi kinerja algoritma. Oleh karena itu, penelitian ini merespon kebutuhan untuk mengisi kesenjangan ini dengan mengoptimalkan algoritma Gaussian Naive Bayes secara khusus untuk prediksi sindrom metabolik, dengan tujuan meningkatkan akurasi prediksi dan relevansi klinisnya dalam konteks penanganan penyakit metabolik.

Penelitian ini bertujuan utama untuk membentuk, menganalisis, dan mengoptimalkan performa algoritma Gaussian Naive Bayes dalam prediksi sindrom metabolik. Dengan menggali potensi algoritma ini, penelitian ini berupaya memberikan kontribusi penting dalam pengembangan model deteksi dini sindrom metabolik. Dengan adanya model yang handal, diharapkan dapat meningkatkan efisiensi diagnosa, mengurangi risiko penyakit, dan secara positif memengaruhi kesehatan masyarakat secara keseluruhan.

2. METODE PENELITIAN

Pada tahap penelitian ini diberikan penjelasan sistematis mengenai urutan proses yang dilakukan dalam penelitian. Tahapan yang diuraikan dalam rangkaian ini dapat dipahami mulai dari analisis kebutuhan hingga hasil penelitian. Penelitian ini melibatkan beberapa tahapan antara lain analisis kebutuhan data, pengumpulan data, *preprocessing* data, pembangunan model menggunakan algoritma *Gaussian Naive Byes* menggunakan bahasa pemrograman Python pada platform Google Colab, evaluasi model menggunakan SMOTE, dan visualisasi hasil. Berdasarkan urutan tersebut maka tahapan penelitian akan digambarkan pada Gambar 1.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.



Gambar 1 Desain Sistem

2.1 *Data Collection*

Tabel 1 Data Sindrom Metabolik

	seqn	Age	Sex	Marital	Income	Race	…	BC	HDL	Tc	MS
0	62161	22	Male	Single	8200.0	White	…	92	41	84	0
1	62164	44	Female	Married	4500.0	White	…	82	28	56	0
2	62169	21	Male	Single	800.0	Asian	…	107	43	78	0
3	62199	57	Male	NaN	9000.0	White	…	100	35	98	1
4	62218	38	Female	Single	8200.0	Black	…	102	36	162	1

Tabel 2 Deskripsi Kolom

Kolom	Deskripsi
seqn	Nomor identifikasi berurutan.
Age	Usia individu.
Sex	Jenis kelamin individu (misalnya, Pria, Wanita).
Marital	Status perkawinan individu.
Income	Tingkat pendapatan atau informasi terkait pendapatan.
Race	Latar belakang etnis atau ras individu.
WaistCirc	Pengukuran lingkar pinggang.
BMI	Indeks Massa Tubuh, ukuran komposisi tubuh.
Albumunuria	Pengukuran terkait albumin dalam urin.
UrAlbCr	Rasio albumin terhadap kreatinin urin.
UricAcid	Kadar asam urat dalam darah.
BC	Kadar glukosa darah, indikator risiko diabetes.
HDL	Kadar kolesterol High-Density Lipoprotein (kolesterol "baik").
Tc	Kadar trigliserida dalam darah.
MS	Variabel biner menunjukkan ada (1) atau tidak adanya (0) sindrom metabolik.

Data yang digunakan diperoleh dari situs Kaggle (*kaggle.com*), sebuah *platform* yang menyediakan *dataset* untuk berbagai proyek *data science*. Data yang dimiliki oleh Albert Antony ini berisi informasi tentang individu dengan sindrom metabolik, suatu kondisi medis kompleks

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

yang terkait dengan sekelompok faktor risiko penyakit kardiovaskular dan diabetes tipe 2. Data tersebut meliputi pengukuran demografi, klinis, dan laboratorium, serta ada tidaknya sindrom metabolik. Data ini memiliki panjang 2402 data dengan 15 kolom atribut. Contoh data sindrom metabolik dapat dilihat pada Tabel 1. Adapun data pada Tabel 1 terdapat 15 kolom atribut yang dapat dideskripsikan dalam Tabel 2.

2.2 *Data Preprocessing*

Dalam tahap ini dilakukan beberapa pembersihan dan penyesuaian terhadap beberapa komponen di dalamnya seperti adanya nilai *null*, *encode* label dll. Tahap ini meliputi *Null-Handling* yang melibatkan penanganan nilai-nilai yang hilang (*null*) dalam *dataset*. *Null-handling* bisa mencakup penghapusan baris atau kolom yang mengandung nilai *null*, atau penggantian nilai *null* dengan nilai yang sesuai, seperti nilai rata-rata atau median. Pada data yang dimiliki terdapat kolom yang bersifat *Null*. Dalam proses ini dilakukan penghapusan baris dalam kolom yang terdapat nilai kosong.

Tahap selanjutnya yaitu *Label-Encoding* yang mana jika terdapat variabel kategori yang bersifat nominal atau ordinal, perlu dilakukan *label encoding* yang dapat dilihat pada Gambar 2. *Label encoding* mengubah nilai-nilai kategori menjadi angka-angka agar dapat diproses oleh algoritma *machine learning*. Tahap ini menggunakan LabelEncoder dari scikit-learn, kolom kategorikal seperti 'Marital', 'Sex', dan 'Race' diubah menjadi nilai numerik. Setelah itu, dilakukan pembuatan salinan *data frame* dengan menghapus kolom target 'MetabolicSyndrome'.

	seqn	Age	Sex	Marital	Income	Race	WaistCirc	BMI	Albuminuria	UrAlbCr	UricAcid	BloodGlucose	HDL	Triglycerides	MetabolicSyndrome
0	62161	22	Male	Single	8200.0	White	81.0	23.3	0	3.88	4.9	92	41	84	0
1	62164	44	Female	Married	4500.0	White	80.1	23.2	0	8.55	4.5	82	28	56	0
2	62169	21	Male	Single	800.0	Asian	69.6	20.1	0	5.07	5.4	107	43	78	0
3	62172	43	Female	Single	2000.0	Black	120.4	33.3	0	5.22	5.0	104	73	141	0
4	62177	51	Male	Married	NaN	Asian	81.1	20.1	0	8.13	5.0	95	43	126	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2396	71901	48	Female	Married	1000.0	Other	NaN	59.7	0	22.11	5.8	152	57	107	0
2397	71904	30	Female	Single	2000.0	Asian	NaN	18.0	0	2.90	7.9	91	90	91	0
2398	71909	28	Male	Single	800.0	MexAmerican	100.8	29.4	0	2.78	6.2	99	47	84	0
2399	71911	27	Male	Married	8200.0	MexAmerican	106.6	31.3	0	4.15	6.2	100	41	124	1
2400	71915	60	Male	Single	6200.0	White	106.6	27.5	0	12.82	5.2	91	36	226	1

2401 rows × 15 columns

Gambar 2 Data Sebelum *Preprocessing*

Sebelum *preprocessing*, terdapat 2401 *record* data yang terlihat pada Gambar 2. Setelah proses tersebut, jumlah data berkurang menjadi 2009 record. *Preprocessing* penting untuk membersihkan dan mempersiapkan data sebelum analisis lebih lanjut.

	seqn	Age	Sex	Marital	Income	Race	WaistCirc	BMI	Albuminuria	UrAlbCr	UricAcid	BloodGlucose	HDL	Triglycerides	MetabolicSyndrome
0	62161	22	1	3	8200.0	5	81.0	23.3	0	3.88	4.9	92	41	84	0
1	62164	44	0	1	4500.0	5	80.1	23.2	0	8.55	4.5	82	28	56	0
2	62169	21	1	3	800.0	0	69.6	20.1	0	5.07	5.4	107	43	78	0
3	62172	43	0	3	2000.0	1	120.4	33.3	0	5.22	5.0	104	73	141	0
5	62178	80	1	4	300.0	5	112.5	28.5	0	9.79	4.8	105	47	100	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2394	71895	31	1	1	2500.0	0	74.0	20.6	0	2.00	6.7	95	64	81	0
2395	71898	65	0	1	5400.0	3	98.5	29.4	0	5.51	6.7	114	49	165	1
2398	71909	28	1	3	800.0	3	100.8	29.4	0	2.78	6.2	99	47	84	0
2399	71911	27	1	1	8200.0	3	106.6	31.3	0	4.15	6.2	100	41	124	1
2400	71915	60	1	3	6200.0	5	106.6	27.5	0	12.82	5.2	91	36	226	1

2009 rows × 15 columns

Gambar 3 Data Setelah *Preprocessing*

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

2.3 *Model Building*

Setelah data sudah bersih dan siap, selanjutnya adalah tahap pembuatan model dengan mengimplementasikan algoritma Gaussian Naive Bayes pada *dataset* yang telah diproses sebelumnya. Terdapat beberapa hal yang dilakukan dalam tahap ini, yang pertama yaitu pembagian data menjadi dua bagian yaitu menjadi data latih (*training data*) dan data uji (*testing data*). Data latih digunakan untuk melatih model, sementara data uji digunakan untuk menguji performa model pada data yang belum pernah dilihat sebelumnya.

Selanjutnya yaitu tahap impelementasi algoritma Gaussian Naive Bayes. Algoritma ini sama seperti teorema bayes di mana persamaannya disajikan pada Pers. (1). Di mana P(A) merupakan nilai probabilitas A, P(B) nilai probabilitas B, P(A|B) nilai probabilitas A menyebabkan B (probabilitas posterior), dan P(B|A) nilai probabilitas B menyebabkan A.

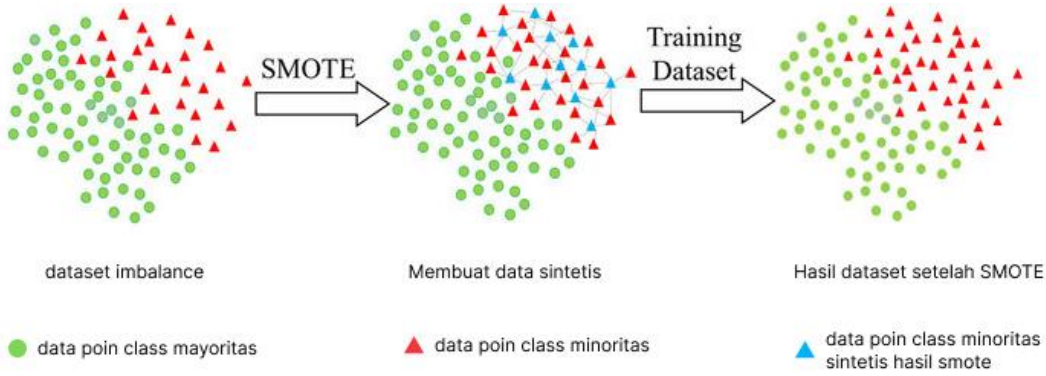
$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{1}$$

Perbedaannya terletak pada perhitungan probabilitas tiap fiturnya menggunakan persamaan distribusi normal seperti yang ditampilkan pada Pers. (2). Di mana P(x(A) | y(B)) merupakan nilai probabilitas fitur x(A) terhadap target y(B), *x* yaitu nilai fitur yang di prediksi, σ nilai Standar deviasi, dan μ merupakan nilai rata-rata.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{2}$$

2.4 *Model Evaluation*

Setelah model berhasil dilatih dan diuji, tahapan selanjutnya melibatkan analisis performa model untuk mengukur sejauh mana keefektifan prediksi. Dalam analisis ini, berbagai metrik evaluasi seperti akurasi, presisi, *recall*, dan F1-*score* dievaluasi sesuai dengan jenis masalah klasifikasi yang dihadapi. Pentingnya memahami metrik-metrik ini adalah untuk mendapatkan wawasan yang komprehensif tentang seberapa baik model dapat mengklasifikasikan data. Selanjutnya, untuk mengatasi ketidakseimbangan kelas dalam *dataset*, dilakukan optimalisasi dengan menerapkan teknik *class balancing* menggunakan SMOTE (Synthetic Minority Over-sampling Technique).



Gambar 4 Visualisasi Cara Kerja SMOTE

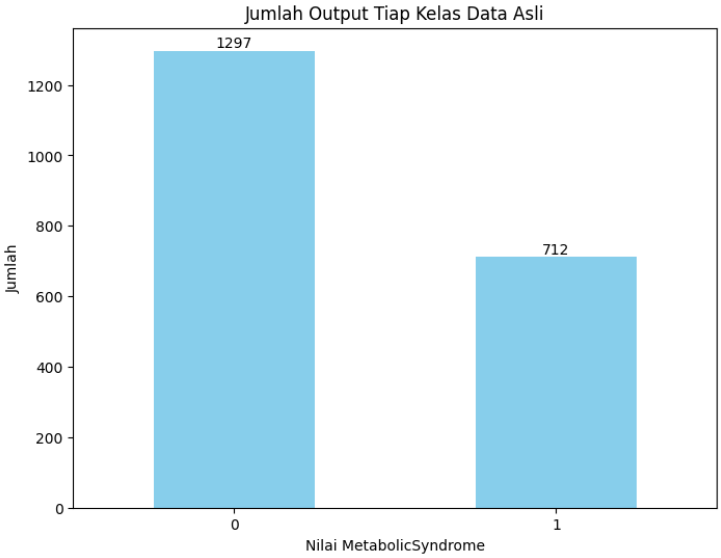
Jika terdapat ketidakseimbangan dalam distribusi kelas pada data, teknik *class balancing* seperti *Synthetic Minority Over-sampling Technique* (SMOTE) dapat digunakan. Cara kerja SMOTE ialah menciptakan sampel-sampel sintetis dari kelas minoritas untuk menyamakan jumlah sampel antara kelas mayoritas dan minoritas seperti yang ditampilkan dalam Gambar 4. Hal ini

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

membantu mencegah model menjadi bias terhadap kelas mayoritas dan meningkatkan kemampuan model dalam mengenali kelas minoritas.

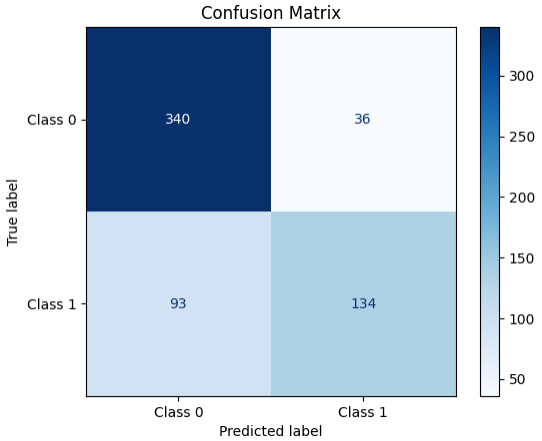
3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, pengujian sistem dilakukan dengan membagi *dataset* menjadi dua bagian utama yaitu 70% data digunakan untuk data *training*, sementara 30% sisanya digunakan untuk menguji performa model atau data *testing*. Pembagian ini bertujuan untuk menghasilkan model klasifikasi yang dapat memahami dan mengklasifikasikan label penyakit *Metabolic Syndrome* menjadi "Ya" atau "Tidak". Pembagian kelas pada data dapat dilihat pada Gambar 5. Terlihat dari Gambar 5, bahwa terjadi ketidakseimbangan kelas pada nilai *Metabolic Syndrome*, sehingga dibutuhkan adanya sebuah prediksi untuk membuktikan apakah data tersebut benar valid atau tidak.



Gambar 5 Pembagian Kelas pada Data *Metabolic Syndrome*

3.1 Hasil Pengujian Menggunakan Gaussian Naïve Bayes



Gambar 6 *Confusion Matrix* Menggunakan Gaussian Naïve Bayes

Pada percobaan pertama yang menggunakan algoritma Gaussian Naive Bayes dapat dilihat pada Gambar 6 yang menghasilkan nilai akurasi, presisi, *recall*, dan F1-*Score* pada Tabel 3.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Setelah dilakukan evaluasi dari model, didapat perbedaan *recall* yang cukup signifikan antara kelas “0” dan kelas “1”. Perbedaan *recall* yang cukup signifikan ini mengindikasikan bahwa model sangat baik dalam menentukan kelas “0” dan kurang baik dalam menentukan kelas “1”. Hal ini terjadi dikarenakan ketidakseimbangan antara data kelas “0” dan “1” pada data pelatihan, sehingga perlu dilakukan penyeimbangan jumlah kelas pada data pelatihan.

Tabel 3 Nilai dari *Confusion Matrix* GNB

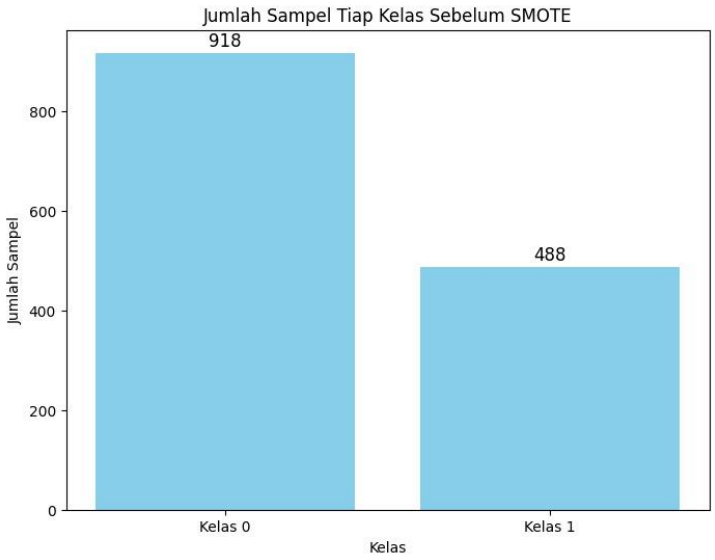
	Precision	***Recall***	***F1-Score***
0	0.81	0.93	0.86
1	0.82	0.59	0.68
Accuracy			**0.81**

3.2 Hasil Pengujian Menambahkan Teknik SMOTE

Salah satu cara untuk mengatasi data kelas yang tidak seimbang adalah dengan dilakukan *oversampling* pada data. Teknik *oversampling* yang umum dipakai adalah SMOTE seperti pada Gambar 4. Adapun proses SMOTE yang digunakan menggunakan *library* ’imblearn.over_sampling import SMOTE’, lalu diinisialisasi dengan mengatur random_state ke nilai tertentu untuk memastikan reproduksibilitas dengan menambahkan *code* ’sm = SMOTE(random_state=2)’. Setelah itu *oversampling* diterapkan dengan penggunaan metode ’*fit_resample’* menjadi fungsi di bawah ini:

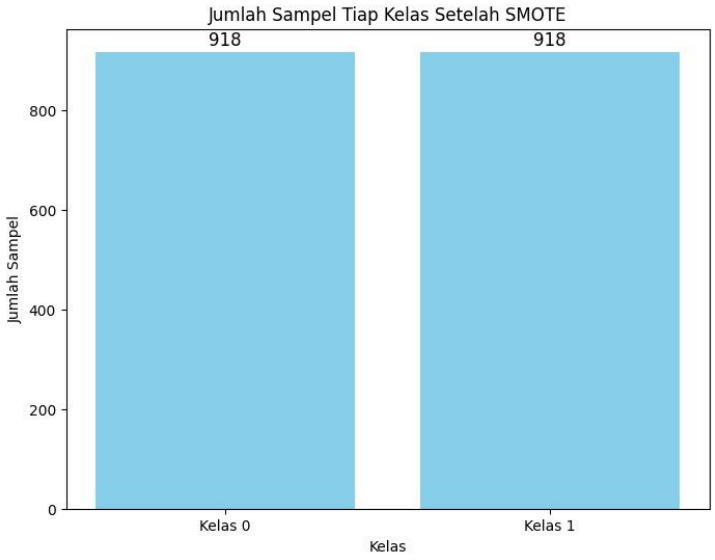
```
x_resampled, y_resampled = sm.fit_resample(x_train, y_train.ravel())
```

Proses ini akan menciptakan contoh sintetis baru dari kelas minoritas (kelas yang jumlahnya lebih sedikit) sehingga jumlah sampel dalam kedua kelas menjadi seimbang. Proses ini membuat *dataset* menjadi lebih seimbang dan membantu model untuk mempelajari pola dari kedua kelas dengan lebih baik, yang diharapkan akan meningkatkan kinerja model dalam melakukan klasifikasi pada kelas minoritas. Adapun perbedaan pembagian kelas dari sebelum dan setelah menggunakan SMOTE dijelaskan pada gambar 7 dan 8. Setelah dilakukan penyeimbangan kelas dengan *oversampling* menggunakan SMOTE dapat dilihat pada Gambar 9 yang menghasilkan nilai akurasi, presisi, *recall*, dan F1-*Score* pada Tabel 4.



Gambar 7 Pembagian Kelas Sebelum SMOTE

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.



Gambar 8 Pembagian Kelas Setelah SMOTE

Gambar 9 *Confusion Matrix* Menggunakan Performa SMOTE

Tabel 4 Nilai dari *Confusion Matrix* Performa SMOTE

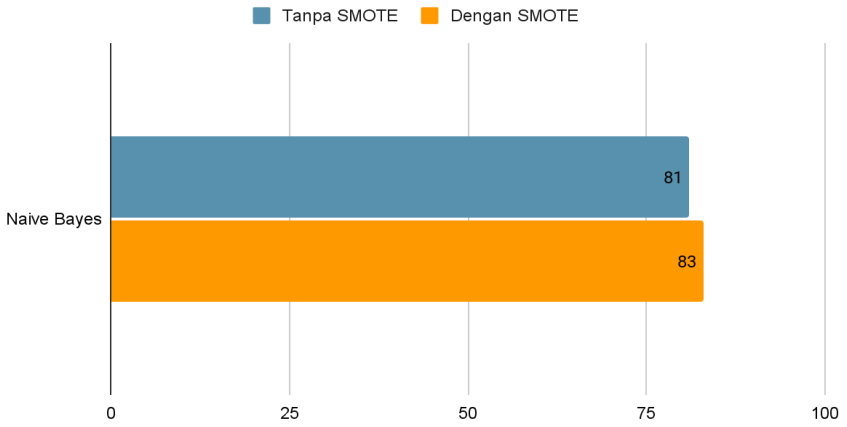
	Precision	***Recall***	***F1-Score***
0	0.86	0.88	0.87
1	0.78	0.73	0.75
Accuracy			**0.83**

Berdasarkan pelatihan model setelah *oversampling* dengan SMOTE, akurasi dari model mengalami kenaikan sebanyak 2% menjadi 83%. Terdapat beberapa perubahan dalam *classification report* kedua. Kenaikan terjadi pada *recall* dan F1-*score* serta terjadi sedikit penurunan pada *precision* terutama pada kelas “1”. Perbandingan akurasi ini juga dapat dilihat pada Gambar 10.

Kenaikan akurasi sebesar 2% dari 81% menjadi 83% pada model kedua yang menggunakan Gaussian Naive Bayes (GNB) setelah penerapan SMOTE mengindikasikan peningkatan yang signifikan dalam kemampuan model untuk mengklasifikasikan kasus sindrom metabolik. SMOTE memberikan kontribusi penting dengan meningkatkan *recall* pada kelas minoritas, yang dapat menjadi kritis dalam aplikasi klinis di mana deteksi yang tepat dari kasus positif sangat penting.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Selain itu, peningkatan F1-*Score* menunjukkan peningkatan keseluruhan dalam kemampuan model untuk melakukan klasifikasi yang akurat. Sampai saat ini belum ada penelitian yang menggunakan teknik SMOTE pada penerapan metode Gaussian Naive Bayes, terlebih lagi pada prediksi Metabolic Syndrome.



Gambar 10 Perbandingan Akurasi

4. KESIMPULAN

Selama pengujian model pertama, diperoleh akurasi sebesar 81%, mengindikasikan kemampuan model dalam mengklasifikasikan sebagian besar kasus secara benar. *Precision* sebesar 88.5% menunjukkan tingkat keakuratan model ketika menyatakan seseorang memiliki sindrom metabolik. *Recall* sebesar 85% menggambarkan kemampuan model untuk menangkap sebagian besar kasus sindrom metabolik yang sebenarnya. F1-*Score* mencapai 86.7%, mencerminkan keseimbangan yang baik antara kemampuan model untuk mengidentifikasi kasus positif dan menghindari *false positive*.

Pada pelatihan model kedua, terjadi peningkatan akurasi menjadi 83%, dengan *recall* pada kelas "1" meningkat dari 0.59 menjadi 0.73. Hal ini menandakan peningkatan signifikan dalam kemampuan model untuk mendeteksi kasus sebenarnya dari kelas "1" dibandingkan dengan model sebelumnya. Meskipun terdapat sedikit penurunan *recall* pada kelas "0", namun hal ini masih dapat diterima. Peningkatan 7% pada F1-*Score* pada kelas "1" dan 1% pada kelas "0" menunjukkan peningkatan performa model dalam mengidentifikasi kasus positif dan menghindari *false positive*.

Hasil ini memberikan dorongan untuk pengembangan lebih lanjut dalam mendekati deteksi Sindrom Metabolik dengan menggunakan metode *machine learning*, dengan potensi peningkatan performa melalui penyesuaian dan peningkatan teknik seperti *oversampling* menggunakan SMOTE. Keseluruhan, penelitian ini memberikan kontribusi pada pemahaman mendalam tentang aplikasi model *machine learning* untuk prediksi Sindrom Metabolik dan memberikan landasan untuk penelitian lebih lanjut dalam pengembangan model yang lebih canggih.

DAFTAR PUSTAKA

Anand, M. V., KiranBala, B., Srividhya, S. R., C., K., Younus, M., & Rahman, M. H. (2022). Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer. *Mobile Information Systems*, *2022*, 1–7. https://doi.org/10.1155/2022/2436946

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Dobrowolski, P., Prejbisz, A., Kuryłowicz, A., Baska, A., Burchardt, P., Chlebus, K., Dzida, G., Jankowski, P., Jaroszewicz, J., Jaworski, P., Kamiński, K., Kapłon-Cieślicka, A., Klocek, M., Kukla, M., Mamcarz, A., Mastalerz-Migas, A., Narkiewicz, K., Ostrowska, L., Śliż, D., … Bogdański, P. (2022). Metabolic syndrome – a new definition and management guidelines A joint position paper by the Polish Society of Hypertension, Polish Society for the Treatment of Obesity, Polish Lipid Association, Polish Association for Study of Liver, Polish Society of Family Medicine, Polish Society of Lifestyle Medicine, Division of Prevention and Epidemiology Polish Cardiac Society, "Club 30" Polish Cardiac Society, and Division of Metabolic and Bariatric Surgery Society of Polish Surgeons. *Archives of Medical Science*, *18*(5), 1133–1156. https://doi.org/10.5114/aoms/152921

Han, T. S., & Lean, M. E. (2016). A clinical perspective of obesity, metabolic syndrome and cardiovascular disease. *JRSM Cardiovascular Disease*, *5*, 204800401663337. https://doi.org/10.1177/2048004016633371

Herningtyas, E. H., & Ng, T. S. (2019). Prevalence and distribution of metabolic syndrome and its components among provinces and ethnic groups in Indonesia. *BMC Public Health*, *19*(1), 1–12. https://doi.org/10.1186/S12889-019-6711-7/FIGURES/3

Hu, X., Li, X.-K., Wen, S., Li, X., Zeng, T.-S., Zhang, J.-Y., Wang, W., Bi, Y., Zhang, Q., Tian, S.-H., Min, J., Wang, Y., Liu, G., Huang, H., Peng, M., Zhang, J., Wu, C., Li, Y.-M., Sun, H., … Chen, L.-L. (2022). Predictive modeling the probability of suffering from metabolic syndrome using machine learning: A population-based study. *Heliyon*, *8*(12), e12343. https://doi.org/10.1016/j.heliyon.2022.e12343

Huang, P. L. (2009). A comprehensive definition for metabolic syndrome. *Disease Models & Mechanisms*, *2*(5–6), 231–237. https://doi.org/10.1242/dmm.001180

Libnao, M., Misula, M., Andres, C., Mariñas, J., & Fabregas, A. (2023). Traffic incident prediction and classification system using naïve bayes algorithm. *Procedia Computer Science*, *227*, 316–325. https://doi.org/10.1016/j.procs.2023.10.530

Palaniappan, L. P., Wong, E. C., Shin, J. J., Fortmann, S. P., & Lauderdale, D. S. (2011). Asian Americans have greater prevalence of metabolic syndrome despite lower body mass index. *International Journal of Obesity*, *35*(3), 393–400. https://doi.org/10.1038/ijo.2010.152

Rochlani, Y., Pothineni, N. V., Kovelamudi, S., & Mehta, J. L. (2017). Metabolic syndrome: Pathophysiology, management, and modulation by natural compounds. *Therapeutic Advances in Cardiovascular Disease*, *11*(8), 215–225. https://doi.org/10.1177/1753944717711379/ASSET/IMAGES/LARGE/10.1177_1753944717711379-FIG1.JPEG

Saklayen, M. G. (2018). The Global Epidemic of the Metabolic Syndrome. *Current Hypertension Reports*, *20*(2), 1–8. https://doi.org/10.1007/S11906-018-0812-Z/METRICS

Tavares, L. D., Manoel, A., Donato, T. H. R., Cesena, F., Minanni, C. A., Kashiwagi, N. M., da Silva, L. P., Amaro, E., & Szlejf, C. (2022). Prediction of metabolic syndrome: A machine learning approach to help primary prevention. *Diabetes Research and Clinical Practice*, *191*, 110047. https://doi.org/10.1016/j.diabres.2022.110047

The GBD 2015 Obesity Collaborators. (2017). Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *New England Journal of Medicine*, *377*(1), 13–27. https://doi.org/10.1056/NEJMOA1614362/SUPPL_FILE/NEJMOA1614362_DISCLOSURES.PDF

Venkata, P., & Pandya, V. (2022). Data mining model and Gaussian Naive Bayes based fault diagnostic analysis of modern power system networks. *Materials Today: Proceedings*, *62*(P13), 7156–7161. https://doi.org/10.1016/j.matpr.2022.03.035

Wilkinson, M. J., Manoogian, E. N. C., Zadourian, A., Lo, H., Fakhouri, S., Shoghi, A., Wang, X., Fleischer, J. G., Navlakha, S., Panda, S., & Taub, P. R. (2020). Ten-Hour Time-Restricted Eating Reduces Weight, Blood Pressure, and Atherogenic Lipids in Patients with Metabolic Syndrome. *Cell Metabolism*, *31*(1), 92-104.e5. https://doi.org/10.1016/j.cmet.2019.11.004

World Health Organization. (2020). *The top 10 causes of death*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Zhou, Y., Wu, T., Jiang, Y., Li, Y., Li, K., Quan, L., & Lyu, Q. (2022). DeepNup: Prediction of Nucleosome Positioning from DNA Sequences Using Deep Neural Network. *Genes*, *13*(11), 1983. https://doi.org/10.3390/genes13111983

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024, Pp. 123 – 133
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Ensemble Learning pada Kategorisasi Produk *E-Commerce* Menggunakan Teknik *Boosting*

Genta Dwigi Sepbriant [(1)*], Danang Wahyu Utomo [(2)]
Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang
e-mail : 111202013101@mhs.dinus.ac.id, danang.wu@dsn.dinus.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 31 Januari 2024, direvisi 16 Maret 2024, diterima 17 Maret 2024, dan dipublikasikan 25 Mei 2024.

Abstract

The development of e-commerce significantly contributes to technological advancement, especially for businesses adopting the concept. The growth of e-commerce has seen a significant increase, reaching 196.47 million users in 2023. In e-commerce, a wide range of product variations is provided to users, which can lead to errors or confusion in product selection. Product categorization is crucial in e-commerce to assist users in navigating efficiently. However, manual categorization is less effective as it can be time-consuming. This study aims to clarify the factors of concern in grouping using the K-Nearest Neighbors (KNN) algorithm in product categorization on the e-commerce platform. This research focuses on whether the novelty lies in the implemented algorithm, the variables used, or the applied grouping parameters. This work applies the XGBoost algorithm to improve the effectiveness of product categorization in e-commerce through ensemble learning approaches. The research findings indicate that boosting algorithms like XGBoost outperform individual algorithms like KNN regarding classification accuracy. This proves that ensemble learning approaches may greatly enhance product classification in e-commerce. The testing process of the implemented e-commerce system in this study also provides confidence in the theoretical and practical benefits of applying this research to enhance efficiency and user experience in product categorization on the e-commerce platform.

Keywords: Product Categorization, E-Commerce, Ensemble Learning, XGBoost, Boosting

Abstrak

Perkembangan *e-commerce* memberikan kontribusi nyata dalam perkembangan tektnologi terutama pada perusahaan yang menjalankan konsep bisnis. Perkembangan *e-commerce* saat ini meningkat secara signifikan dengam mencapai 196,47 juta jiwa pada tahun 2023. Dalam *e-commerce* tentunya banyak memberikan variasi produk pilihan kepada pengguna. Hal ini dapat mengakibatkan kesalahan atau kekeliruan dalam pemilihan produk. Kategorisasi produk menjadi penting dalam sebuah *e-commerce* untuk membantu pengguna menavigasi dengan efisien. Namun, kategorisasi manual kurang efektif karena dapat memakan waktu yang cukup lama. Penelitian ini bertujuan untuk mengklarifikasi faktor-faktor yang menjadi perhatian dalam pengelompokan menggunakan algoritma K-Nearest Neighbors (KNN) pada kategorisasi produk dalam platform *e-commerce*. Fokus penelitian ini adalah pada apakah keunikan (*novelty*) terletak pada algoritma yang diimplementasikan, variabel yang digunakan, atau pada parameter pengelompokkan yang diterapkan. Dengan menggunakan teknik *ensemble learning*, penelitian ini mengimplementasikan algoritma XGBoost untuk meningkatkan efisiensi kategorisasi produk dalam *e-commerce*. Hasil penelitian menunjukkan bahwa algoritma *boosting* seperti XGBoost mampu mengungguli kinerja algoritma individu seperti KNN dalam hal akurasi klasifikasi. Ini menunjukkan bahwa dengan memanfaatkan teknik *ensemble learning*, kategorisasi produk dalam *e-commerce* dapat ditingkatkan secara signifikan. Proses pengujian sistem *e-commerce* yang diimplementasikan dalam penelitian ini turut memberikan keyakinan terkait manfaat penerapan penelitian ini secara teoritis maupun praktis dalam meningkatkan efisiensi dan pengalaman pengguna dalam kategorisasi produk di platform *e-commerce*.

Kata Kunci: Kategorisasi Produk, *E-Commerce*, *Ensemble Learning*, XGBoost, *Boosting*

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

1. PENDAHULUAN

Saat ini *e-commerce* menunjukkan perkembangan yang signifikan dalam layanan jual beli melalui internet. *E-commerce* memberikan kontribusi nyata dalam perkembangan teknologi terutama bagi perusahaan yang menjalankan konsep bisnis. *E-commerce* menjadi isu dalam kehidupan sehari-hari terutama bagi pelaku bisnisatau Perusahaan (Gomero-Fanny et al., 2021). Para pelaku bisnis, organisasi dan/atau Perusahaan saat ini menjalankan layanan *e-commerce* melalui telepon selular yang dapat memudahkan para pengguna dalam hubungan bisnis dan komunikasi. Selain itu, juga memudahkan dalam melakukan jangkauan pelanggan, perluasan jaringan pemasaran dan transaksi dalam bentuk apapun. Perkembangan *e-commerce* saat ini mampu mengubah aturan bisnis dan segala jenis transaksi (Huang et al., 2019). Pada penelitian lainnya juga menyatakan bahwa saat ini Perusahaan menggunakan perkembangan teknologi dalam platform jual beli *online* (Indasari & Tjahyanto, 2023). *E-commerce* menjadi solusi utama dalam bagi pelaku bisnis dalam menawarkan produknya dalam singkat, sedangkan bagi pengguna (*user*) memudahkan pencarian produk dalam waktu singkat.

Dalam era digital saat ini, jumlah data yang besar dan meningkat memberikan tantangan baru dalam pengolahan informasi. *E-commerce* menjadi salah satu isu yang melibatkan transaksi data dalam jumlah besar. Berdasarkan data pada laman dataindonesia.id menunjukkan bahwa pengguna *e-commerce* tahun 2023 meningkat dari tahun sebelumnya yaitu 10% menjadi 196,47 juta jiwa. Dari data tersebut dapat disimpulkan jika pengguna *e-commerce* melakukan transaksi, maka terjadi transaksi yang melibatkan data dan informasi dalam jumlah besar. Adanya transaksi dalam jumlah besar menjadikan perusahaan melakukan investasi besar terhadap platform *e-commerce*. Menurut Mashalah et al. (2022), terdapat 3 (tiga) faktor yang mendorong perkembangan *e-commerce*: teknologi pendukung, persaingan, dan perilaku pengguna.

Permasalahan umum yang banyak dibahas dalam *e-commerce* adalah banyaknya jumlah produk atau jenis produk yang dipasarkan pada platform *online* (Ansharullah et al., 2023; Donati et al., 2019; Indasari & Tjahyanto, 2023). Banyaknya variasi produk memberikan banyak pilihan produk kepada pengguna yang dapat menyebabkan kesalahan atau kekeliruan dalam pemilihan produk. Bagi perusahaan, hal ini menjadi masalah besar karena ada potensi produk tertentu tidak dipilih oleh konsumen. Bagi pengguna yang masih awam tentang platform *e-commerce*, banyaknya variasi produk dapat menyebabkan kesalahan pemilihan produk karena harus memahami deskripsi produk satu per satu. Penelitian lain menyatakan permasalahan bahwa produk yang sama dijual pada beberapa toko, namun dengan informasi yang berbeda (Ristoski et al., 2018). Perlu adanya kategorisasi produk untuk membantu pengguna awam dalam menemukan produk yang sesuai.

Dalam platform *e-commerce* populer, menyediakan banyak ragam produk yang ditampilkan. Masing-masing produk memberikan deskripsi dan ulasan yang berbeda. Dibutuhkan adanya kategorisasi untuk produk-produk tersebut dengan tujuan: rekomendasi produk, prediksi produk (Jain & Kumar, 2020), dan kategorisasi yang sesuai dengan deskripsi (Tan et al., 2020). Kategorisasi secara manual dengan pemberian label pada masing-masing produk membutuhkan waktu lama dan tidak ada jaminan ketepatan dalam penempatan deskripsi produk dalam suatu kategori. Menurut Jahanshahi et al. (2021), mengusulkan kategori yang sesuai dengan deskripsi membutuhkan waktu yang lama dan kompleksitas semakin meningkat.

Permasalahan lainnya, kategorisasi produk dapat dipengaruhi berdasarkan input nama produk. Adanya kesamaan nama produk dengan deskripsi yang berbeda dapat mempengaruhi ketepatan dalam kategorisasi produk. Pada penelitian yang dilakukan Jahanshahi et al. (2021) menunjukan adanya perbedaan kategori dengan sub kategori berdasarkan input nama produk. Menurut Kim et al. (2021), kategorisasi yang dilakukan manusia sangat sulit dilakukan untuk mendapat hasil yang akurat dan cepat hanya berdasarkan nama. Ketidaktepatan kategorisasi produk juga dapat dilakukan oleh operator (Ozyegen et al., 2022). Rekomendasi kategorisasi oleh operator dapat menghasilkan kategorisasi yang subyektif atau kurang tepat. Adanya penambahan atau *update* produk baru dalam *e-commerce* dapat mempengaruhi tingkat akurasi kategorisasi produk.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Berdasarkan permasalahan di atas, beberapa solusi telah diusulkan untuk kategorisasi produk yang lebih baik yaitu: teknik penyematan kata (Sharma & Sagvekar, 2023); klasifikasi (Pawłowski, 2022; Perdana et al., 2021); dan translasi mesin (Tan et al., 2020). Kategorisasi produk berbasis klasifikasi teks menjadi solusi yang populer dan terbukti mampu menempatkan produk *e-commerce* sesuai dengan kategorinya. Dari penelitian yang dilakukan oleh Patra et al. (2021), hasil menunjukkan bahwa klasifikasi produk menggunakan model pembelajaran mesin (*machine learning* atau ML) mampu memberikan akurasi terbaik dalam mengklasifikan produk berdasarkan kategori yang telah ditentukan. Pada penelitian lainnya, Pothuganti (2019) mengusulkan algoritma *Ordered Weighted Averaging Combination* (OWC) untuk mengenali identitas produk baru yang tidak dikenali. Tujuannya, memudahkan kategorisasi produk yang baru ditambahkan pada platform *e-commerce*. Pengklasifikasi produk otomatis diusulkan oleh Lee & Yoon (2018) dengan tujuan untuk mengklasifikasikan produk berdasarkan deskripsi dan dokumen doc2vec. Teknik, model, dan algoritma klasifikasi yang diusulkan masing-masing memiliki performa baik dalam mengklasifikasikan produk *e-commerce*.

Pada pendekatan *data mining*, model *Machine Learning (*ML) telah diusulkan untuk meningkatkan performa, akurasi, dan kinerja model yang diusulkan. *Ensemble learning* adalah pendekatan yang terdiri dari penggabungan model *Machine Learning* (ML) untuk meningkatkan kinerja kategorisasi atau klasifikasi. Arumnisaa & Wijayanto (2023) mengusulkan *ensemble classifier* untuk meningkatkan akurasi dalam klasifikasi. Hasil akurasi yang dihasilkan lebih tinggi dari algoritma individu. Pada penelitian lainnya diusulkan analisis perbandingan terhadap model *ensemble* dengan *neural network* untuk mengetahui akurasi terbaik dalam klasifikasi produk *e-commerce* (Kalaivani, 2020). Pada perbandingan tersebut, algoritma AdaBoost dengan SVM menghasilkan akurasi terbaik. Pada penelitian lainnya, *ensemble learning* diusulkan dalam klasifikasi *review* produk dengan hasil bahwa algoritma *boosting* mampu mengungguli algoritma individu seperti *K-Nearest Neighbors* (KNN) dalam akurasi klasifikasi (Fayaz et al., 2020). Algoritma *boosting* seperti *extreme gradient boosting* (XGBoost), AdaBoost mampu memberikan akurasi terbaik dibandingkan algoritma individu dalam pembelajaran mesin.

Pada penelitian ini mengusulkan algoritma XGBoost dalam kategorisasi produk *e-commerce*. XGBoost merupakan pendekatan berbasis *boosting* yang terdiri dari beberapa *decision tree* di mana pohon sebelum dan berikutnya akan saling bergantungan. XGBoost menggabungkan berbagai metode pengklasifikasian lemah dengan melatih model baru secara berurutan dengan menggunakan model klasifikasi sebelumnya. Pada eksperimen menggunakan data uji *e-commerce dataset* yang terdiri dari atribut deskripsi dan kategori dengan jumlah *dataset* 50.434 data produk. Uji coba juga menerapkan *hyperparameter tunning* untuk mencari akurasi terbaik.

2. METODE PENELITIAN

2.1 *Ensemble Learning*

Ensemble learning adalah metode dalam *Machine Learning* (ML) dengan menggabungkan dua atau lebih algoritma dengan tujuan meningkatkan akurasi, mengurangi kesalahan, atau bias model yang dihasilkan oleh algoritma individu. Mienye juga menyatakan definisi dari *ensemble learning* yaitu suatu teknik dengan kombinasi algoritma *Machine Learning* (ML) untuk mendapatkan performa superior dibandingkan dengan algoritma yang digunakan secara tunggal (Mienye & Sun, 2022). Peneliti lain menyatakan bahwa *ensemble learning* merupakan metode yang melibatkan beberapa algoritma dengan performa lemah kemudian dikombinasikan untuk mengasilkan performa yang lebih baik (Dong et al., 2020). Secara umum, tujuan utama dari metode *ensemble learning* adalah menggabungkan algoritma ML untuk mencari performa terbaik dari penggabungan yang dilakukan. Menurut Mienye, terdapat 3 (tiga) teknik yaitu: *bagging, boosting,* dan *stacking.* Penelitian ini fokus pada teknik *boosting* yaitu teknik dengan pemrosesan sekuensial berdasarkan proses dari model sebelumnya.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

2.2 Pendekatan *Boosting*

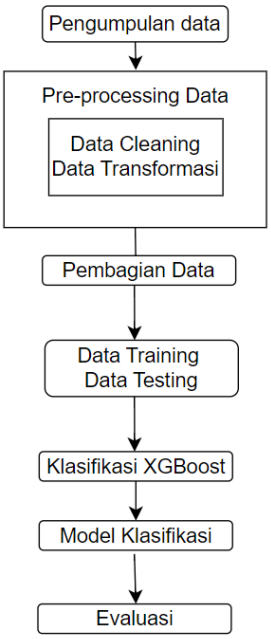
Pendekatan *boosting* adalah metode yang melibatkan penerapan secara *iterative* dari algoritma *boosting* dasar ke versi yang disesuaikan dari data masukan (Mienye & Sun, 2022). Metode *boosting* biasanya menggunakan data masukan untuk melatih model *boosting* lemah, kesalahan pada klasifikasi, dan melatih model tersebut dengan set yang disesuaikan pada klasifikasi sebelumnya. Peneliti lain menyatakan bahwa metode *boosting* ini berfokus pada pengurangan bias daripada variasi dengan meningkatkan *boosting* awal dasar yang memiliki bias tinggi. Secara keseluruhan pendekatan *boosting* adalah metode yang digunakan untuk mengurangi kesalahan dalam analisis data prediktif dan meningkatkan kinerja model pada data kompleks atau sulit diklasifikasikan dan dapat bekerja baik dengan berbagai algoritma.

2.3 Algoritma XGBoost

Algoritma XGBoost merupakan algoritma dengan implementasi *decision-tree* yang menggunakan kerangka dari *Gradient Boosting*. XGBoost merupakan implementasi dari *Gradient Boosting* untuk meningkatkan performa kinerja dan stabilitas. Pada kasus klasifikasi dan regresi, algoritma XGBoost tepat diusulkan karena mengacu pada pohon keputusan terbaik. Menurut Jafarzadeh et al menyatakan bahwa algoritma XGBoost adalah algoritma terbaik jika dibandingkan dengan algoritma ML lainnya (Jafarzadeh et al., 2021). Beberapa keuntungan algoritma XGBoost adalah mampu menangani *dataset* yang besar, data yang hilang, dan mencegah terjadinya *overfitting* (Nobre & Neves, 2019). Adanya penggunaan *tree depth, learning rate* dan *subsampling* menjadikan algoritma XGBoost menghasilkan akurasi lebih baik dibandingkan algoritma *Machine Learning* (ML) lainnya. Adapum formula dan variabel yang digunakan dalam koteks penggunaan algoritma XGBoost untuk klasifikasi dalam *e-commerce* dirumuskan pada Pers. (1). Di mana $F(x)$ adalah fungsi prediksi untuk kelas tertentu, ω_0 adalah bias, M adalah jumlah pohon keputusan (*boosting* rounds), dan $f_m(x)$ adalah fungsi dari pohon keputusan ke-m.

$$F(x) = \omega_0 + \sum_{m=1}^{M} f_m(x) \quad (1)$$

2.4 *Dataset*



Gambar 1 Tahap Eksperimen

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Pada eksperimen menggunakan *e-commerce dataset* sebagai uji kategorisasi produk. Penulis melakukan pengambilan data dari *website* https://www.kaggle.com/datasets/saurabhshahane/ecommerce-text-classification. *Dataset* ini merupakan data untuk klasifikasi berdasarkan katagori pada platfrom *e-commerce*, yang terdiri dari 4 kategori utama, yaitu "*Electronics", "Household", "Books"*, dan *"Clothing & Accessories"* yang hampir mencakup 80% dari produk yang biasanya ada disitus *e-commerce*. *Dataset* ini disajikan dalam format ".csv" dengan kolom pertama adalah nama kategori dan kolom kedua data poin (deskripsi produk) dari kategori tersebut. pengaturan atribut adalah *label* dan *desc*.

Pengaturan *dataset* menggunakan alat bantu Google Colaboratory. Pada kolom *label* merupakan kolom kategori dari suatu produk *e-commerce* yaitu: *household, books, clothing & accessories,* dan *electronics*. Kolom *desc* sebagai kolom deskripsi suatu produk yaitu berisi penjelasan informasi produk. *Dataset* menampilkan sampel *dataset* dengan pengaturan atribut adalah *label* dan *desc*. Pengaturan *dataset* menggunakan alat bantu Google Colaboratory. Pada kolom *label* merupakan kolom kategori dari suatu produk *e-commerce* yaitu: *household, books, clothing & accessories,* dan *electronics*. Kolom *desc* sebagai kolom deskripsi suatu produk yaitu berisi penjelasan informasi produk.

2.5 Eksperimen

Tahap eksperimen sesuai dengan tahapan pada Gambar 1. Tahapannya yaitu terdiri dari pengumpulan data, pemrosesan awal data untuk melakukan pembersihan *dataset*, transformasi data, pengaturan kategorisasi produk, dan evaluasi menggunakan *confusion matrix.* Eksperimen menggunakan alat bantu Google Colaboratory dengan bahasa pemrograman Python.

2.5.1 Pemrosesan Awal Data

Pemrosesan awal data digunakan untuk melakukan normalisasi data seperti tahap pembersihan, transformasi data, dan integrasi data untuk disiapkan sebelum digunakan pada tahap analisis. Berikut adalah fungsi yang diterapkan pada pemrosesan awal data:

1) Konversi huruf dalam ukuran kecil atau *lowercase*.
2) Menghilangkan tanda baca seperti titik, koma, tanda tanya, petik satu, petik dua, dan strip.
3) *Stopword removal* menghilangkan kata yang tidak relevan berdasarkan daftar *library* yang ditentukan.
4) Menghilangkan teks numerik.

2.5.2 Tokenisasi

Tokenasi merupakan sebuah proses untuk membagi sejumlah teks baik kalimat maupun paragraf menjadi beberapa bagian tertentu. Dengan kata lain tokenisasi adalah tahapan untuk memotong struktur kalimat menjadi perkata. Fungsi tokenisasi dijalankan pada urutan ke 5 (lima) setelah menghilangkan teks numerik. Sebagai contoh kalimat "*Pitaara Box Offer Exclus collage"* kemudian diubah menjadi *"Pitaara"*, *"Box"*, *"Offer"*, *"Exclus"*, dan *"collage"*.

2.5.3 Lemmatisasi dan *Stemming*

Lemmatisasi adalah proses pengelompokan bentuk infleksi yang berbeda dari sebuah kata sehingga dapat dianalisis sebagai satu *item*. Lemmatisasi mirip dengan *stemming* dengan membawa konteks kata-kata dengan menggabungkan kata-kata yang mirip dengan satu kata. Lemmatisasi contoh kalimat "*textured print which gives*" menjadi "*texture print which give"* sedangkan untuk *stemming* contoh kalimat "*Paper plane design framed wall hanging motivationa*l" menjadi "*Paper plane design frame wall hang motiv*".

2.5.4 Transformasi Data

Berdasarkan jenis *dataset* yang digunakan, *dataset* tidak dapat langsung digunakan pada XGBoost karena *dataset* sendiri berupa data tekstual. Sebaliknya, diperlukan langkah transformasi data dari bentuk teks ke bentuk numerik menggunakan TF-IDF (Jahanshahi et al.,

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

2021). TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan gabungan dari dua konsep yaitu menghitung bobot dan nilai dengan menghitung frekuensi yang muncul pada teks berupa kata dokumen (Andriani & Wibowo, 2021). Tujuan dari metode ini yaitu untuk menentukan bobot pada kata-kata dalam teks dengan mengukur tingkat signifikansi suatu kata dalam kumpulan dokumen. Rumus perhitungan TF-IDF dituliskan pada Pers. (2) sampai (4), di mana $Tf_{i\,j}$ merupakan banyaknya kata i pada dokumen ke j, IDf_i yaitu banyaknya dokumen yang mengandung kata ke i, N adalah total dokumen, dan log merupakan logaritma natural.

$$Tf_{ij} = \frac{f_{i,j}}{\sum k\, f_{i,j}} \tag{2}$$

$$IDf_i = log\left(\frac{N}{n_i}\right) + 1 \tag{3}$$

$$TFIDF_{ij} = TF_{ij} \times IDf_i \tag{4}$$

2.5.5 Skenario Uji

Pada kategorisasi produk uji *dataset* dilakukan pengaturan pembagian data latih dan data uji. Besaran persentase data uji adalah 20% atau 0,2 dari ukuran *dataset*. Pengaturan selanjutnya adalah *hyperparameter tuning* yaitu pengaturan parameter pada implementasi fungsi XGBoost berdasarkan parameter *number of tree* atau *n_estimators, learning_rate,* dan *max_depth*. Parameter *n_estimators* digunakan untuk menentukan jumlah pohon keputusan yang dibuat secara paralel, *learning rate* adalah parameter yang digunakan sebagai laju pembelajaran dan *max_depth* adalah parameter yang digunakan untuk penentu kedalaman maksimum. Pada eksperimen ini dilakukan uji coba dengan pengaturan: *n_estimator*s=[100, 200]; *max_depth*=[3, 5, 7]; dan *learning_rate*=[0.1, 0.2]. Teknik *random search* diusulkan untuk mendapatkan hasil optimal pada kombinasi *hyperparameter* antara *n_estimators, max_depth* dan *learning_rate.*

2.5.6 *Confusion Matrix*

Confusion matrix digunakan sebagai alat bantu untuk evaluasi klasifikasi atau kategorisasi produk *e-commerce*. Kesesuaian atau ketetapan deskripsi produk pada suatu kategori diukur dengan *confusion matrix.* Pada Gambar 2, *confusion matrix* berupa tabel yang terdiri dari jumlah data kelas positif dan prediksi benar (TP), kelas negatif prediksi benar (FP), kelas positif prediksi salah (FN) dan kelas negatif prediksi salah (TN).

	Actual positive	***Actual negative***
Predicted positive	TP	FP
Predicted negative	FN	TN

Gambar 2 *Confusion Matrix*

Berdasarkan hasil uji akurasi setelah mengklasifikasikan produk, dilakukan dua kali pengujian seperti yang ditunjukkan pada Tabel 1. Pengujian pertama yakni dengan nilai *n_estimator:* 100 dan *n_estimator:* 200. Hasil dari pengujian pertama menunjukan bahwa *n_estimator:* 100, *learning_rate:* 0,2 dan *max_depth: 7* diperoleh akurasi sebesar 93,41%. Sedangkan pengujian kedua dengan nilai *n_estimator:* 200, pengujian dilakukan *n_estimator:* 200, *learning_rate:* 0,2 dan *max_depth: 7* diperoleh akurasi tertinggi sebesar 94,17%. Berdasarkan hasil tersebut, menunjukan bahwa pengujian yang dilakukan menemukan hasil yang mampu melakukan klasifikasi dengan baik.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024: 123 – 133

Tabel 1 Hasil Uji Akurasi

Hyperparameter			Akurasi
n_estimators	learning_rate	max_depth	
100	0.1	3	86.85%
		5	90.99%
		7	91.83%
	0.2	3	91.35%
		5	92.77%
		7	93.41%
200	0.1	3	91.47%
		5	92.71%
		7	93.41%
	0.2	3	92.86%
		5	93.79%
		7	94.17%

3. HASIL DAN PEMBAHASAN

3.1 *Preprocessing*

Pada tahap pertama eksperimen dimulai dengan persiapan *dataset* kemudian dilakukan pembersihan data. Dari proses normalisasi data menghasilkan data yang berbeda dengan *dataset* sebelumnya. Pada Gambar 3, menunjukkan sampel *dataset* setelah dilakukan normalisasi data. Sampel data menunjukkan deskripsi suatu produk bersih tidak ada tanda baca atau simbol tertentu. Kemudian tulisan dalam bentuk *lowercase* dan kata dalam bentuk kata dasar (setelah melalui proses lemmatisasi dan *stemming*). Hal ini ditujukan untuk memudahkan komputasi pada model XGBoost dalam kategorisasi produk. Intensitas kemunculan masing-masing kata menjadi penentu kategori untuk deskripsi tersebut.

```
'pitaara box romant venic canva paint 6mm thick mdf frame 21 1 14  enclosur materi mdf mount frame 1 14  5
3 6 35 6  size 21 1 14 0  53 6 35 6  enhanc beauti room wall breathtak digit print artwork print technolog
captur everi detail imag print enhanc matt paint canva ensur rich live colour wall art panel mount mdf rea
di hang wall beauti interior home artwork gift live dine room outdoor galleri hotel restaur offic recept k
itchen area balconi bathroom pitaara box offer exclus collect thousand artwork digit paint canva print wal
l poster wall decor product home offic surround provid rang creativ spectacular art product use gift everi
occas everi season tag wall paint canva print modern art abstract design wallart artwork home bedroom dine
live draw room digit print bathroom common area kitchen offic decor stretch stretch frame frame beauti cla
ssi royal special uniqu eleg stylish creativ afford best photo gift fabric balconi interior exterior outdo
or galleri hotel restaur colour color small larg extra larg overs hang giant slim durabl waterproof buy sh
op purchas decor onlin place vintag canva romant venic artwork paint style'
```

Gambar 3 Sampel Data *Preprocessing*

Tahap selanjutnya adalah lemmatisasi dan *stemming*, yaitu mengubah kata atau kalimat menjadi kata dasar, sebagai contoh ditunjukkan pada Gambar 4. Proses *stemming* menjadikan input tiap kata ke bentuk akar atau dasar. Beberapa kata dasar mungkin asing atau dikenali karena ada beberapa yang dipotong. Sebagai contoh kata *romantic* setelah dilakukan stemming menjadi *romant*. Proses selanjutnya adalah transformasi data dengan melakukan TF-IDF yaitu melakukan vektorisasi teks mengubah teks dalam representasi vektor numerik. Hasil dari vektorisasi TF-IDF diterapkan pada model yang diusulkan untuk dihitung nilai akurasinya.

Input: Pitaara Box Romantic Venice Canvas Painting
Output: pitaara box romant venic canva

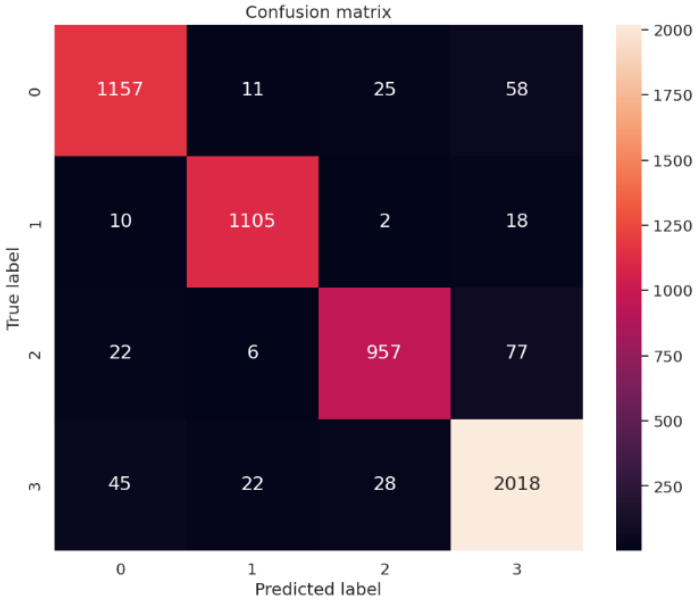
Gambar 4 Contoh Proses *Stemming*

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

3.2 *Confusion Matrix*

Proses selanjutnya adalah confusion matrik dengan mengevaluasi kinerja model klasifikasi dengan menyajikan informasi tentang hasil prediksi model terhadap data uji. Dari Gambar 5 menunjukkan *confusion matrix* memiliki empat sel utama, yang masing-masing menyajikan informasi yaitu:

1) *True Positive* (TP): Terletak pada diagonal utama matriks, di mana prediksi dan aktual keduanya positif.
2) *True Negative* (TN): Terletak di luar diagonal utama, di antara sel-sel yang tidak terlibat dalam *True Positive*.
3) *False Positive* (FP): Terletak di kolom yang sesuai dengan prediksi positif tetapi aktualnya negatif. Misalnya, untuk Prediksi 0, FN/FP terletak di baris Aktual 1, 2, dan 3.
4) *False Negative* (FN): Terletak di baris yang sesuai dengan aktual positif tetapi diprediksi sebagai negatif. Misalnya, untuk Aktual 0, FN/FP terletak di kolom Prediksi 1, 2, dan 3.



Gambar 5 Hasil *Confusion Matrix*

Tabel 2 Metrik Evaluasi

	Precision	Recall	F1-Score
0	0.92	0.94	0.93
1	0.97	0.97	0.97
2	0.90	0.95	0.92
3	0.96	0.93	0.94
	Akurasi		**0.94**

3.2.1 *Precision*

Presisi digunakan untuk mengukur seberapa akurat prediksi positif dalam kategorisasi produk. Dalam kasus kategorisasi produk ini, presisi digunakan mengetahui keakuratan produk yang diprediksi masuk dalam kategorisasi yang benar. Pada Tabel 2 menunjukkan bahwa produk yang dikategorikan pada label 0 adalah 94%, label 1 96%, label 2 95%, dan label 3 90%. Hal ini menunjukkan bahwa di atas 90% produk yang dikategorisasikan tepat atau benar, atau dengan kata lain di atas 90% model yang diprediksi positif yaitu benar-benar positif.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

3.2.2 *Recall*

Recall digunakan untuk mengetahui nilai prediksi positif sebenarnya dalam kategorisasi produk. Sebagai contoh pada Tabel 2, label 1, *recall* menunjukkan 0,97 atau 97%. Artinya, 97% dari hasil positif sesuai diprediksi dengan tepat menggunakan algoritma XGBoost. Dalam hal ini, semakin tinggi persentase *recall* maka algoritma XGBoost semakin baik dalam menangani kategorisasi produk.

3.2.3 *F1-Score*

F1-Score merupakan matriks evaluasi yang digunakan untuk mencari keseimbangan antara nilai hasil *precision* dan *recall*. Pada Tabel 2, label 1 menunjukkan *F1-Score* adalah 0,97 atau 97% hal ini menunjukkan ada keseimbangan antara *precision* dan *recall*. Hal ini juga menunjukkan bahwa algoritma XGBoost mampu meminimalkan nilai *False Positive* dan *False Negative*.

4. KESIMPULAN

Setelah melakukan analisis dengan mengumpulkan data dan melakukan pengujian, penelitian ini mengusulkan penerapan XGboost dalam kategorisasi produk di platform *e-commerce* sebagai solusi umtuk meningkatkan akurasi klasifikasi. Dalam eksperimen yang dilakukan melalui pengujian dan variasi *hyperparameter*, ditemukan bahwa konfigurasi seperti *n_estimator*: 200, *learning_rate*: 0,2, dan *max_depth*: 7 mampu menghasilkan akurasi tertinggi sebesar 97,17%. Penggunaan matrik evaluasi seperti *precision*, *recall*, dan *F1-score* dalam mengevaluasi kinerja model menunjukan bahwa algoritma XGboost mampu memberikan prediksi yang akurat untuk setiap kategori produk, menawarkan potensi untuk meningkatkan efisiensi bisnis serta pengalaman pengguna dalam memilih produk dengan lebih efisien di lingkungan *e-commerce*. Dengan mengandalkan teknik *ensemble learning*, XGBoost dapat mengatasi tantangan dalam klasifikasi produk, mengoptimalkan proses kategorisasi, dan secara signifikan meningkatkan akurasi, yang pada gilirannya akan memberikan dampak positif pada operasional dan keuntungan bisnis. Kesimpulannya, penelitian ini memberikan wawasan yang berharga tentang penerapan XGBoost dalam konteks *e-commerce*, menyoroti potensi algoritma ini sebagai solusi efektif untuk meningkatkan kualitas layanan dan pengalaman pengguna di platform perdagangan *online*.

DAFTAR PUSTAKA

Andriani, N., & Wibowo, A. (2021). Implementasi Text Mining Klasifikasi Topik Tugas Akhir Mahasiswa Teknik Informatika Menggunakan Pembobotan TF-IDF dan Metode Cosine Similarity Berbasis Web. *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer Dan Aplikasinya*, *2*(2), 130–137. https://conference.upnvj.ac.id/index.php/senamika/article/view/1807

Ansharullah, M. O., Agustin, W., Lusiana, Junadhi, Erlinda, S., & Zoromi, F. (2023). Product Classification Based on Categories and Customer Interests on the Shopee Marketplace Using the Naïve Bayes Method. *JAIA - Journal of Artificial Intelligence and Applications*, *2*(2), 15–22. https://doi.org/10.33372/jaia.v2i2.888

Arumnisaa, R. I., & Wijayanto, A. W. (2023). Comparison of Ensemble Learning Method: Random Forest, Support Vector Machine, AdaBoost for Classification Human Development Index (HDI). *SISTEMASI*, *12*(1), 206. https://doi.org/10.32520/stmsi.v12i1.2501

Donati, L., Iotti, E., Mordonini, G., & Prati, A. (2019). Fashion Product Classification through Deep Learning and Computer Vision. *Applied Sciences*, *9*(7), 1385. https://doi.org/10.3390/app9071385

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258. https://doi.org/10.1007/S11704-019-8208-Z/METRICS

Fayaz, M., Khan, A., Rahman, J. U., Alharbi, A., Uddin, M. I., & Alouffi, B. (2020). Ensemble Machine Learning Model for Classification of Spam Product Reviews. *Complexity*, *2020*, 1–10. https://doi.org/10.1155/2020/8857570

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Gomero-Fanny, V., Ruiz, A., & Andrade-Arenas, L. (2021). Prototype of Web System for Organizations Dedicated to e-Commerce under the SCRUM Methodology. *International Journal of Advanced Computer Science and Applications*, *12*(1), 437–444. https://doi.org/10.14569/IJACSA.2021.0120152

Huang, Y., Chai, Y., Liu, Y., & Shen, J. (2019). Architecture of next-generation e-commerce platform. *Tsinghua Science and Technology*, *24*(1), 18–29. https://doi.org/10.26599/TST.2018.9010067

Indasari, S. S., & Tjahyanto, A. (2023). Automatic Categorization of Multi Marketplace FMCGs Products using TF-IDF and PCA Features. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *12*(2), 198–204. https://doi.org/10.32736/sisfokom.v12i2.1621

Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F., & Homayouni, S. (2021). Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation. *Remote Sensing*, *13*(21), 4405. https://doi.org/10.3390/rs13214405

Jahanshahi, H., Ozyegen, O., Cevik, M., Bulut, B., Yigit, D., Gonen, F. F., & Başar, A. (2021). *Text Classification for Predicting Multi-level Product Categories*. http://arxiv.org/abs/2109.01084

Jain, S., & Kumar, V. (2020). Garment Categorization Using Data Mining Techniques. *Symmetry*, *12*(6), 984. https://doi.org/10.3390/sym12060984

Kalaivani, P. (2020). Machine Learning Approach to Analyse Ensemble Models and Neural Network Model for E-Commerce Application. *Indian Journal of Science and Technology*, *13*(28), 2849–2857. https://doi.org/10.17485/IJST/v13i28.927

Kim, H., Joo, G., & Im, H. (2021). Product Category Classification using Word Embedding and GRUs. *The Journal of Korean Institute of Information Technology*, *19*(4), 11–18. https://doi.org/10.14801/jkiit.2021.19.4.11

Lee, H., & Yoon, Y. (2018). Engineering doc2vec for automatic classification of product descriptions on O2O applications. *Electronic Commerce Research*, *18*(3), 433–456. https://doi.org/10.1007/S10660-017-9268-5/METRICS

Mashalah, H. Al, Hassini, E., Gunasekaran, A., & Bhatt (Mishra), D. (2022). The impact of digital transformation on supply chains through e-commerce: Literature review and a conceptual framework. *Transportation Research Part E: Logistics and Transportation Review*, *165*, 102837. https://doi.org/10.1016/j.tre.2022.102837

Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Nobre, J., & Neves, R. F. (2019). Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets. *Expert Systems with Applications*, *125*, 181–194. https://doi.org/10.1016/j.eswa.2019.01.083

Ozyegen, O., Jahanshahi, H., Cevik, M., Bulut, B., Yigit, D., Gonen, F. F., & Başar, A. (2022). Classifying multi-level product categories using dynamic masking and transformer models. *Journal of Data, Information and Management*, *4*(1), 71–85. https://doi.org/10.1007/s42488-022-00066-6

Patra, A., Vivek, V., Shambhavi, B. R., Sindhu, K., & Balaji, S. (2021). Product Classification in E-Commerce Sites. In *Advances in Intelligent Systems and Computing: Vol. 1299 AISC* (pp. 485–495). Springer, Singapore. https://doi.org/10.1007/978-981-33-4299-6_40

Pawłowski, M. (2022). Machine Learning Based Product Classification for eCommerce. *Journal of Computer Information Systems*, *62*(4), 730–739. https://doi.org/10.1080/08874417.2021.1910880

Perdana, S. A. P., Aji, T. B., & Ferdiana, R. (2021). Aspect Category Classification dengan Pendekatan Machine Learning Menggunakan Dataset Bahasa Indonesia. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *10*(3), 229–235. https://doi.org/10.22146/jnteti.v10i3.1819

Pothuganti, K. (2019). Open-World Classification Algorithm to Product Identification. *International Journal of Innovative Research in Computer and Communication Engineering*, *7*(12), 4282–4287. https://doi.org/10.2139/ssrn.3719055

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Ristoski, P., Petrovski, P., Mika, P., & Paulheim, H. (2018). A machine learning approach for product matching and categorization. *Semantic Web*, *9*(5), 707–728. https://doi.org/10.3233/SW-180300

Sharma, P., & Sagvekar, V. R. (2023). Weighted Ensemble LSTM Model with Word Embedding Attention for E-Commerce Product Recommendation. *Journal of Communications Software and Systems*, *19*(4), 299–307. https://doi.org/10.24138/jcomss-2023-0126

Tan, L., Li, M. Y., & Kok, S. (2020). E-Commerce Product Categorization via Machine Translation. *ACM Transactions on Management Information Systems*, *11*(3), 1–14. https://doi.org/10.1145/3382189

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024, Pp. 134 – 146
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Klasterisasi Jumlah Penduduk Provinsi Jawa Timur Tahun 2021-2023 Menggunakan Algoritma K-Means

Risqi Pradana Aryanto [(1)*], Agung Nilogiri [(2)], Ari Eko Wardoyo [(3)]
Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember, Jember
e-mail : riskipradana221001@gmail.com, {agungnilogiri,arieko}@unmuhjember.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 9 Februari 2024, direvisi 16 Maret 2024, diterima 17 Maret 2024, dan dipublikasikan 25 Mei 2024.

Abstract

Understanding the population data of a region is crucial for policy development and strategic planning. East Java Province, the second-largest province in Indonesia, has undergone significant population growth from 2021 to 2023. Uneven growth poses challenges in resource and infrastructure management. The K-Means algorithm clusters population data into several groups based on specific characteristics. The Elbow *method is used to determine the optimal number of clusters, ensuring the accuracy of the analysis. This research aims to analyze and cluster the population distribution in each city in East Java Province, providing a more detailed and accurate depiction. The research findings reveal three significant clusters. Cluster 0 includes 21 towns, Cluster 1 comprises 4, and Cluster 2 encompasses 13. These findings have important implications for targeted development policy formulation at the city level in East Java Province. Additionally, this study contributes to the development of demographic analysis and population management, using valid methods and consistent results between RapidMiner and manual calculations. In conclusion, this research provides a solid foundation for more effective development policy formulation in East Java Province, offering essential information for sustainable population management.*

Keywords: Population Distribution, Clustering, East Java, K-Means, Elbow Method, Data Mining

Abstrak

Pemahaman terhadap data populasi suatu wilayah menjadi hal yang sangat penting untuk pengembangan kebijakan dan perencanaan strategis. Provinsi Jawa Timur, sebagai provinsi kedua terbesar di Indonesia, mengalami perkembangan penduduk signifikan dari tahun 2021 hingga 2023. Pertumbuhan yang tidak merata menimbulkan tantangan dalam pengelolaan sumber daya dan infrastruktur. Algoritma K-Means digunakan sebagai solusi untuk mengelompokkan data jumlah penduduk ke dalam beberapa kelompok berdasarkan karakteristik tertentu. Metode Elbow digunakan untuk menentukan jumlah klaster optimal, memastikan keakuratan analisis. Tujuan penelitian ini adalah menganalisis dan mengklasterisasi sebaran penduduk di setiap kota di Provinsi Jawa Timur, memberikan gambaran yang lebih rinci dan akurat. Hasil penelitian menunjukkan adanya tiga klaster yang signifikan. *Cluster* 0 mencakup 21 kota, *Cluster* 1 mencakup 4 kota, dan *Cluster* 2 mencakup 13 kota. Temuan ini memiliki implikasi penting untuk perumusan kebijakan pembangunan yang lebih tepat sasaran di tingkat kota di Provinsi Jawa Timur. Selain itu, penelitian ini memberikan kontribusi pada pengembangan bidang analisis demografis dan pengelolaan populasi, dengan metode yang valid dan hasil yang konsisten antara penggunaan RapidMiner dan perhitungan manual. Kesimpulannya, penelitian ini menyediakan landasan yang kuat bagi perumusan kebijakan pembangunan yang lebih efektif di Provinsi Jawa Timur, memberikan informasi yang diperlukan untuk pengelolaan populasi yang berkelanjutan.

Kata Kunci: Sebaran Penduduk, Pengelompokan, Jawa Timur, K-Means, Metode Elbow, *Data Mining*

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

1. PENDAHULUAN

Pada era modern ini, pemahaman yang mendalam terhadap data populasi suatu wilayah menjadi sangat krusial untuk pengembangan kebijakan dan perencanaan strategis. Provinsi Jawa Timur, yang terletak di bagian timur Pulau Jawa, Indonesia, adalah provinsi dengan jumlah penduduk terbesar kedua di negara ini. Provinsi Jawa Timur merupakan salah satu wilayah yang terus mengalami perkembangan penduduk yang signifikan dari tahun 2021 hingga 2023. Fenomena ini menjadi sorotan karena adanya kebutuhan untuk memahami dan mengelola distribusi penduduk secara efektif. Dengan lebih dari 40 juta jiwa, Jawa Timur menjadi rumah bagi berbagai kelompok etnis, budaya, dan ekonomi.

Pertumbuhan penduduk di Jawa Timur memiliki pola yang unik dan beragam. Beberapa wilayah mengalami pertumbuhan penduduk yang pesat, sementara wilayah lainnya mengalami pertumbuhan yang lebih lambat atau bahkan penurunan jumlah penduduk. Faktor-faktor seperti urbanisasi, migrasi, dan perubahan sosial-ekonomi berkontribusi terhadap dinamika ini. Pertumbuhan penduduk yang tidak merata di berbagai kota di Provinsi Jawa Timur menimbulkan tantangan tersendiri dalam mengelola sumber daya dan pembangunan infrastruktur. Oleh karena itu, diperlukan suatu pendekatan analisis yang dapat mengidentifikasi pola dan karakteristik dalam distribusi jumlah penduduk.

Meskipun data jumlah penduduk tersedia, masih terdapat kendala dalam memahami pola sebaran penduduk pada tingkat kota di Provinsi Jawa Timur. Ketidakmampuan dalam mengidentifikasi kelompok atau pola tertentu dapat menghambat upaya perencanaan pembangunan yang efisien. Oleh karena itu, perlu dilakukan analisis lebih lanjut untuk mengatasi masalah ini. Maka digunakan algoritma K-Means sebagai metode analisis untuk mengklasterisasi jumlah penduduk pada setiap kota di Provinsi Jawa Timur. K-Means merupakan algoritma klasterisasi yang dapat mengelompokkan data ke dalam beberapa kelompok berdasarkan kesamaan karakteristik tertentu (Nofiar et al., 2019). Penggunaan algoritma ini diharapkan dapat memberikan gambaran yang jelas mengenai pola sebaran penduduk di wilayah tersebut.

K-Means bekerja dengan mengelompokkan data ke dalam k kelompok (klaster) yang ditentukan sebelumnya, dengan meminimalkan variasi atau jarak antara titik data dalam satu klaster. Pendekatan ini sangat relevan dalam konteks analisis sebaran penduduk, karena dapat membantu mengidentifikasi pola sebaran yang mungkin tidak terlihat secara langsung (Talakua et al., 2017). Dengan menggunakan K-Means, dapat mengelompokkan kota-kota berdasarkan tingkat persebaran jumlah penduduk dari rentang tahun 2021 hingga tahun 2023. Selain itu, untuk menentukan jumlah klaster yang optimal, penelitian ini akan menerapkan metode siku (Elbow *method*). Metode ini melibatkan pengujian berbagai jumlah klaster dan mengamati tingkat variasi yang dijelaskan oleh model terhadap jumlah klaster tersebut.

Metode Elbow atau Elbow *method* merupakan suatu teknik yang dimanfaatkan untuk mengidentifikasi jumlah klaster yang optimal dalam proses klastering, terutama diterapkan pada algoritma K-Means. Konsep inti dari metode ini adalah dengan mengamati penurunan varians di dalam setiap klaster seiring dengan variasi jumlah klaster yang berbeda (Syahfitri et al., 2023). Metode Elbow bekerja dengan mengukur varians atau dispersi data dalam setiap klaster saat jumlah klaster berubah-ubah. Proses ini melibatkan iterasi dengan berbagai jumlah klaster, dan pada setiap iterasi, varians dihitung untuk setiap klaster. Hasilnya kemudian dianalisis untuk mengidentifikasi titik di mana penurunan varians menjadi kurang signifikan, menandakan bahwa penambahan klaster tidak lagi memberikan keuntungan substansial dalam mengurangi dispersi data (Fahrozi et al., 2023). Dengan menggunakan metode Elbow, peneliti mengambil keputusan yang lebih informasional dan berbasis data dalam menentukan jumlah klaster optimal untuk suatu *dataset* tertentu. Teknik ini menjadi kunci dalam memahami struktur data dan mengoptimalkan kinerja algoritma klastering, khususnya algoritma K-Means, untuk mencapai segmentasi yang optimal dan lebih bermakna.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Penelitian terkait dengan penggunaan K-Means antara lain “Penggunaan K-Means untuk Klasterisasi Penetapan Instruktur Diklat pada PT PLN (Persero) UDIKLAT Jakarta” (Budiana et al., 2019), “Penggunaan K-Means *Clustering* untuk Klasterisasi Tingkat Kehadiran Dosen” (Virgo et al., 2020), dan “Penggunaan K-Means untuk Klastering Sayuran Unggulan” (Harahap et al., 2022). Namun pada penelitian Harahap tidak digunakan metode lain untuk menentukan jumlah klaster atau k sehingga klaster yang dihasilkan kurang begitu optimal. Penelitian lainnya yang berkaitan dengan penggunaan K-Means digunakan untuk klasterisasi siswa yang berprestasi (Dewi et al., 2022). Pada penelitian ini, jumlah kluster ditentukan menggunakan metode Davies Bouldin sehingga dapat menghasilkan jumlah kluster yang optimal. Penelitian lainnya terkait penggunaan metode Elbow yaitu metode ini digunakan untuk *clustering* pemerataan bantuan sosial di Kabupaten Bojonegoro (Fitriyah et al., 2023) dan algoritma K-Means *Clustering* Metode Elbow digunakan untuk menganalisa motivasi pengunjung Festival Halal JHF (Wicaksana et al., 2023).

Tujuan utama dari penelitian ini adalah untuk menganalisis dan mengklasterisasi sebaran jumlah penduduk pada setiap kota di Provinsi Jawa Timur menggunakan algoritma K-Means. Metode K-Means akan diaplikasikan dengan menggunakan metode Elbow untuk menentukan jumlah klaster yang optimal. Maka dari itu, penelitian ini bertujuan untuk mendapatkan informasi yang lebih rinci dan akurat mengenai pola sebaran penduduk di tingkat kota. Diharapkan penelitian ini bisa berkontribusi terhadap pengembangan bidang analisis demografis dan pengelolaan populasi serta menjadi landasan yang kuat bagi perumusan kebijakan pembangunan yang lebih efektif di Provinsi Jawa Timur.

2. METODE PENELITIAN

Metodologi penelitian merupakan landasan yang digunakan oleh peneliti untuk melaksanakan penelitian. Landasan ini mencakup serangkaian langkah dalam pengelolaan data, dimulai dari analisis kebutuhan hingga pemahaman terhadap hasil penelitian. Proses tersebut dijelaskan melalui beberapa tahapan yang terstruktur dan sistematis, sebagaimana yang ditampilkan dalam Gambar 1.

Gambar 1 Alur Kerja Penelitian

Berdasarkan Gambar 1 yang menggambarkan alur kerja penelitian, penelitian ini dimulai dengan pengumpulan data dan mengolah data yang akan menjadi data utama. Selanjutnya, dilakukan proses analisa data yang melibatkan tahap pembersihan data, dan transformasi data untuk menggabungkan serta mengubah format data agar sesuai dengan kebutuhan. Tahap selanjutnya, yaitu *data mining*, pada tahap ini, peneliti menerapkan metode K-Means *clustering* untuk implementasi *data mining*. Langkah terakhir melibatkan pengujian hasil, yang dilakukan dengan perhitungan manual dan menggunakan aplikasi RapidMiner.

2.1 Pengumpulan Data

Pengumpulan data merupakan suatu proses untuk mendapatkan informasi atau data yang diperlukan untuk menjawab pertanyaan dari penelitian yang diajukan. Proses ini melibatkan kegiatan sistematis dalam mencari dan menghimpun data secara langsung dari sumbernya baik data di lapangan maupun data dari sumber literatur lainnya (Anufia & Alhamid, 2019). Tujuan utama dari pengumpulan data adalah untuk memperoleh informasi yang relevan dan fakta yang dapat memberikan pemahaman yang mendalam terhadap permasalahan penelitian yang tengah diteliti.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

2.2 Mengolah Data

Pengolahan data merupakan serangkaian kegiatan, yaitu pengumpulan, pemrosesan, dan analisis data. Proses pengolahan data bertujuan untuk menghasilkan informasi yang akan digunakan dalam penelitian. Hasil dari proses ini dapat berupa laporan, grafik, atau tabel yang memberikan representasi visual atau naratif dari temuan dan hasil analisis (Nawassyarif et al., 2020).

2.3 Menganalisa Data

Menganalisa data adalah langkah yang dilakukan untuk mengubah data yang telah dikumpulkan dan dibersihkan menjadi informasi yang memiliki nilai dan bermanfaat. Tujuan utama dari analisis data adalah untuk mengidentifikasi tren, pola, serta hubungan yang dapat ditemukan antara berbagai data set yang berbeda. Melalui proses ini, peneliti mampu mengidentifikasi dan menemukan karakteristik dari data yang telah terkumpul (Herviany et al., 2021).

2.4 *Data Mining*

Data mining merupakan suatu proses yang dilakukan untuk menemukan informasi atau pola menarik dalam *dataset* yang telah dipilih. Dalam melaksanakan proses ini, digunakan teknik atau metode khusus yang dapat mencakup berbagai algoritma. Keberhasilan dalam mencapai tujuan dan keseluruhan proses Penambangan Data Pengetahuan (KDD) sangat tergantung pada pemilihan metode atau algoritma yang tepat. Hal ini karena setiap metode memiliki dampak yang signifikan terhadap hasil akhir dan interpretasi data yang dihasilkan. Oleh karena itu, kebijakan yang cermat dalam memilih metode atau algoritma menjadi kunci kesuksesan dalam mengoptimalkan potensi informasi yang dapat ditemukan melalui *data mining* (Naldy & Andri, 2021).

2.5 K-Means

K-Means adalah suatu algoritma pengelompokan yang beroperasi secara iteratif dengan melakukan partisi untuk mengklasifikasikan atau mengelompokkan sejumlah besar objek. Algoritma ini secara berulang melakukan proses pengelompokan dengan membagi objek-objek tersebut ke dalam kluster atau kelompok yang memiliki kesamaan berdasarkan karakteristik tertentu (Triandini et al., 2021). K-Means merupakan salah satu metode *clustering* non-hirarki yang berupaya mempartisi data yang ada ke dalam satu atau lebih kelompok atau kluster. Dalam konteks ini, metode non-hirarki mengacu pada pendekatan di mana kluster tidak memiliki tingkatan atau struktur hirarkis dan objek-objek data dikelompokkan berdasarkan kemiripan mereka ke dalam *cluster* tertentu. Tujuan utama dari K-Means adalah membentuk kelompok-kelompok yang saling homogen dan meminimalkan variasi antara objek-objek dalam satu kelompok dengan objek-objek dalam kelompok lainnya (Triyansyah & Fitrianah, 2018).
Tahapan dari proses K-Means meliputi (Virgo et al., 2020):

1) Inputkan data yang akan dilakukan pengklasteran.
2) Tentukan jumlah kluster yang diinginkan.
3) Tentukan pusat kluster atau *centroid* awal.
4) Lakukan perhitungan jarak Euclidean, proses pengklasteran data, dan perhitungan *centroid* baru untuk iterasi ke-n.
5) Tentukan hasil akhir dari proses pengklasteran.

2.6 *Clustering*

Clustering adalah suatu proses pengelompokan atau pembagian data dalam suatu himpunan menjadi beberapa kelompok, di mana kesamaan data dalam satu kelompok lebih besar dibandingkan dengan kesamaan data tersebut dengan kelompok lainnya. Potensi dari teknik *clustering* ini dapat digunakan untuk mengidentifikasi struktur dalam data, yang nantinya dapat diterapkan dalam berbagai aplikasi seperti klasifikasi, pengolahan gambar, dan pengenalan pola.

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Teknik *clustering* memiliki dua metode pengelompokan, yaitu *hierarchical clustering* dan *non-hierarchical clustering*. *Hierarchical clustering* adalah metode pengelompokan data yang mengelompokkan dua data atau lebih yang memiliki kesamaan atau kemiripan. Proses ini dilanjutkan dengan mengelompokkan objek lain yang memiliki kedekatan dua dan proses ini terus berlangsung hingga terbentuk suatu struktur pohon (*tree*) dengan tingkatan hirarki yang jelas antar objek, mulai dari yang paling mirip hingga yang paling tidak mirip. Meskipun secara logika, pada akhirnya, semua objek akan membentuk sebuah *cluster* (Sadewo et al., 2018).

2.7 Elbow *Method*

Metode Elbow adalah teknik yang digunakan dalam analisis data dan pembelajaran mesin untuk menentukan jumlah klaster optimal dalam suatu *dataset*. Metode ini melibatkan plot variasi yang dijelaskan oleh jumlah klaster yang berbeda dan mengidentifikasi titik "Elbow" atau "siku", di mana tingkat variasi menurun tajam dan stabil, menunjukkan jumlah klaster yang tepat untuk analisis atau pelatihan model (Sholeh et al., 2022).

Berikut adalah langkah-langkah metode Elbow dalam klasterisasi K-means (Yudhistira & Andika, 2023):

1) Pilih jumlah klaster untuk *dataset* (K).
2) Pilih k *centroid* secara acak dari *dataset*.
3) Gunakan jarak Euclidean sebagai metrik untuk menghitung jarak titik dari *centroid* terdekat dan menetapkan titik ke *centroid* klaster terdekat, sehingga menciptakan k klaster.
4) Kemudian *centroid* baru dari klaster yang terbentuk.
5) Tetaplam seluruh titik data berdasarkan *centroid* baru ini, lalu ulangi langkah 4.
6) Lanjutkan langkah ini untuk sejumlah iterasi yang diberikan sampai posisi *centroid* tidak berubah, yaitu tidak ada lagi konvergensi.

2.8 RapidMiner

RapidMiner merupakan sebuah platform perangkat lunak ilmu data yang telah dikembangkan oleh perusahaan bernama RapidMiner. Platform ini menyediakan lingkungan terintegrasi yang mencakup berbagai fungsi seperti persiapan data, pembelajaran mesin, pembelajaran dalam, penambangan teks, dan analisis prediktif. Dengan kata lain, RapidMiner menyediakan berbagai alat dan fitur dalam satu kesatuan platform untuk mendukung sejumlah besar kegiatan di bidang ilmu data, mulai dari pengolahan data hingga pengembangan model prediktif. Platform ini dirancang untuk memfasilitasi tugas-tugas analisis data yang kompleks dan memungkinkan para pengguna untuk mengintegrasikan berbagai aspek dari siklus hidup analisis data dalam satu lingkungan yang terpusat (Anjelita et al., 2019).

3. HASIL DAN PEMBAHASAN

3.1 Sumber Data

Sumber data pada penelitian ini yaitu jumlah penduduk dari setiap kota di provinsi Jawa Timur dari tahun 2021-2023 (Badan Pusat Statistik, 2024). Data di dapat dari *website* https://kedirikota.bps.go.id/ yang merupakan sumber data utama karena menyediakan informasi yang akurat dan terverifikasi mengenai statistik kependudukan. Pemilihan situs ini sebagai sumber utama didasarkan pada reputasinya dalam menyajikan data resmi yang dapat diandalkan. Selain itu, penelitian ini juga didukung oleh berbagai literatur terkait dengan metode *clustering* dan penerapan algoritma K-Means. Literatur ini digunakan sebagai acuan teoritis untuk memperkuat metodologi yang digunakan dalam analisis data.

3.2 Menyeleksi Data

Data yang digunakan dalam penelitian ini berasal dari sumber yang dapat dipercaya, yaitu data dari *website* https://kedirikota.bps.go.id/. Data ini mencakup informasi jumlah penduduk di setiap kota di Provinsi Jawa Timur untuk periode tahun 2021 hingga 2023. Pemilihan data yang valid

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

dan representatif menjadi kunci dalam memastikan keakuratan analisis klasterisasi. Data yang tidak lengkap atau tidak akurat dapat menghasilkan hasil yang bias atau tidak dapat diandalkan. Setelah melalui proses seleksi, *dataset* yang dihasilkan mencakup variabel jumlah penduduk untuk setiap kota, menciptakan data yang valid untuk analisis klasterisasi menggunakan algoritma K-Means. Hasil seleksi data ini akan membentuk landasan yang kuat untuk memahami pola sebaran penduduk di Provinsi Jawa Timur dan membantu mencapai tujuan penelitian yang telah ditetapkan.

3.3 Mengolah Data

Data dari *website* kemudian diolah untuk mendapatkan data yang dapat dianalisis menggunakan algoritma K-Means dengan mudah. Data yang telah diolah akan menjadi *dataset* yang nantinya akan ditentukan klasternya menggunakan algoritma K-Means. Data yang telah di olah terlihat pada Tabel 1.

Tabel 1 *Dataset* Jumlah Penduduk Setiap Kota di Jawa Timur dengan Rentang 2021-2023

Wilayah	2021	2022	2023
Pacitan	589.108	592.916	596.649
Ponorogo	955.839	964.253	972.582
Trenggalek	734.888	739.669	744.358
Tulungagung	1.096.588	1.105.337	1.113.973
Blitar	1.231.013	1.240.322	1.249.497
Kediri	1.644.400	1.656.020	1.667.450
Malang	2.668.296	2.685.900	2.703.175
Lumajang	1.127.094	1.137.227	1.147.261
Jember	2.550.360	2.567.718	2.584.771
Banyuwangi	1.718.462	1.731.731	1.744.814
Bondowoso	778.525	781.417	784.192
Situbondo	688.337	691.260	694.081
Probolinggo	1.155.894	1.159.965	1.163.859
Pasuruan	1.611.805	1.619.035	1.626.029
Sidoarjo	2.091.930	2.103.401	2.114.588
Mojokerto	1.125.522	1.133.584	1.141.516
Jombang	1.325.914	1.335.972	1.345.886
Nganjuk	1.109.683	1.117.033	1.124.247
Madiun	750.143	757.665	765.135
Magetan	674.133	678.343	682.466
Ngawi	873.346	877.432	881.393
Bojonegoro	1.307.602	1.315.125	1.322.474
Tuban	1.203.127	1.209.543	1.215.795
Lamongan	1.356.027	1.371.509	1.386.941
Gresik	1.320.570	1.332.664	1.344.648
Bangkalan	1.071.712	1.086.620	1.101.556
Sampang	976.020	984.162	992.210
Pamekasan	853.507	857.818	862.009
Sumenep	1.129.822	1.136.632	1.143.295
Kota Kediri	287.962	289.418	290.836
Kota Blitar	150.371	151.960	153.541
Kota Malang	844.933	846.126	847.182
Kota Probolinggo	241.202	243.200	245.174
Kota Pasuruan	209.528	211.497	213.450
Kota Mojokerto	133.272	134.350	135.414
Kota Madiun	196.917	199.192	201.460
Kota Surabaya	2.880.284	2.887.223	2.893.698
Kota Batu	214.653	216.735	218.802

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

3.4 Implementasi K-Means

3.4.1 Menentukan Jumlah k Berdasarkan Elbow *Method*

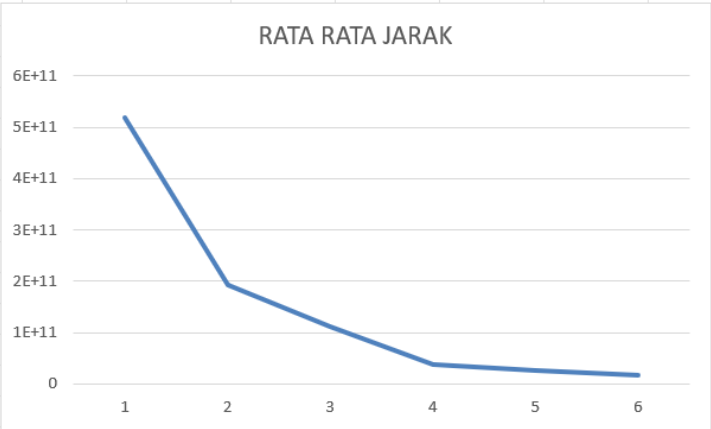
Metode Elbow dipilih dengan cara mengukur performa *distance centroid* menggunakan aplikasi RapidMiner dengan berbagai variasi kluster. Metode Elbow dipilih karena merupakan salah satu metode yang umum digunakan dan relatif mudah dipahami dalam menentukan jumlah klaster yang optimal. Metode Elbow digunakan untuk mengidentifikasi titik di mana penurunan varian antara jumlah klaster yang berbeda menjadi kurang signifikan. Keunggulan metode Elbow adalah kemampuannya untuk memberikan panduan yang jelas dalam menentukan jumlah klaster yang sesuai dengan *dataset* tertentu tanpa memerlukan asumsi sebelumnya tentang distribusi data.

Langkah pertama penentuan jumlah kluster optimal menggunakan metode Elbow adalah berbagai jumlah klaster dipilih dan algoritma K-Means diterapkan untuk masing-masing jumlah klaster tersebut. Kemudian, variasi yang dijelaskan oleh model terhadap jumlah klaster dievaluasi. Grafik yang menunjukkan hubungan antara jumlah klaster dan variasi ini digunakan untuk mengidentifikasi titik "Elbow" atau "siku", yaitu titik di mana penurunan variasi mulai menurun secara signifikan. Pada penelitian ini, titik "siku" terletak pada jumlah klaster tertentu, yang kemudian dipilih sebagai jumlah klaster optimal.

Hasil dari analisis banyak kluster dengan rata-rata jarak dapat ditemukan pada Tabel 2. Berdasarkan grafik metode Elbow pada Gambar 2, terlihat bahwa garis siku terletak di angka 2 pada saat k berjumlah 3 dikarenakan k dimulai dari 0. Sehingga dapat ditentukan jumlah klaster yaitu sebanyak 3 klaster dengan kategori klaster C0, C1, dan C2 dengan kategori masing-masing *cluster* dapat dilihat pada Tabel 3.

Tabel 2 Perhitungan Jarak *Cluster* Setiap Banyak *Cluster*

Banyak *Cluster*	Rata-Rata Jarak
2	5.17948E+11
3	1.93276E+11
4	1.12487E+11
5	38470766310
6	27295170984
7	16188264246



Gambar 2 Grafik Elbow *Method*

Tabel 3 Kategori Klaster

C0	Kota dengan sebaran penduduk sedang
C1	Kota dengan sebaran penduduk terbesar
C2	Kota dengan sebaran penduduk terkecil

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

3.4.2 Menentukan Pusat Klaster

Berdasarkan data jumlah penduduk dari Tabel 1, maka dapat diambil tiga contoh data sebagai pusat klaster atau pusat *centroid* di mana data pada C0 diambil dari banyak penduduk pada Tuban, data pada C1 diambil dari banyak penduduk dari Jember, dan C2 diambil dari banyak penduduk dari kota Kediri. Contoh data pusat klaster dapat dilihat pada Tabel 4. Proses penentuan pusat klaster ini dilakukan untuk memahami karakteristik masing-masing klaster dan dapat memberikan tinjauan yang lebih jelas terhadap pola sebaran penduduk di Provinsi Jawa Timur.

Tabel 4 Pusat Klaster/*Centroid*

Centroid	2021	2022	2023
C0	1.203.127	1.209.543	1.215.795
C1	2.550.360	2.567.718	2.584.771
C2	287.962	289.418	290.836

3.4.3 Menentukan Euclidean *Distance*

$$D_{(i,j)} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \cdots + (x_{ki} - x_{kj})^2} \quad (1)$$

Dalam menghitung Euclidean *distance* seperti pada Pers. (1), jarak antara data ke-i dan pusat *cluster* ke-j dilambangkan sebagai $D(i,j)$. Jarak ini dihitung untuk menentukan seberapa dekat data ke-i berada dengan pusat *cluster* ke-j. Selain itu, x_{ki} merujuk pada nilai data ke-i pada atribut data ke-k. Ini berarti bahwa setiap data memiliki beberapa atribut atau fitur, dan x_{ki} menunjukkan nilai spesifik dari atribut ke-k untuk data ke-i tersebut. Selanjutnya, x_{kj} digunakan untuk menunjukkan titik pusat *cluster* ke-j pada atribut ke-k. Titik pusat ini adalah nilai rata-rata dari semua data dalam *cluster* tersebut untuk atribut ke-k, yang digunakan sebagai referensi untuk menghitung jarak antara data dan pusat *cluster*. Selanjutnya menentukan Euclidean *distance* berdasarkan *dataset* dengan *centroid* yang ditentukan.

$$D\,(1,1) = \sqrt{(589108 - 1203127)^2} + (592916 - 1209543)^2 + (596649 - 1215795)^2 = 1067984$$

$$D\,(1,2) = \sqrt{(589108 - 2550360)^2} + (592916 - 2567718)^2 + (596649 - 2584771)^2 = 3420377$$

$$D\,(1,3) = \sqrt{(589108 - 287962)^2} + (592916 - 289418)^2 + (596649 - 290836)^2 = 525662$$

Dalam proses pengambilan sampel yang melibatkan tiga langkah, perhitungan jarak antara data ke-1 dan ke-3 terhadap *centroid* data dapat dilakukan dengan menggunakan fungsi SQRT yang tersedia dalam perangkat lunak Microsoft Excel. Selanjutnya, klaster untuk setiap data dapat ditentukan berdasarkan tiga klaster dengan jarak terdekat dari setiap data, atau klaster mana yang memberikan nilai jarak Euclidean terkecil untuk setiap data. Informasi tentang penentuan klaster berdasarkan nilai jarak Euclidean dapat ditemukan dalam Tabel 5.

3.4.4 Menentukan *Centroid* Baru

Dalam tahap menentukan *centroid* Baru, setelah klaster untuk setiap data telah ditentukan, langkah berikutnya adalah menghitung nilai rata-rata dari setiap kolom data pada klaster yang sama. Proses ini dilakukan dengan cara menjumlahkan semua data pada kolom yang sesuai dalam setiap klaster, kemudian hasil penjumlahan tersebut dibagi oleh jumlah data pada kolom tersebut. Misalnya, untuk klaster C0, jumlah semua nilai atribut dari setiap data dalam klaster tersebut dijumlahkan, kemudian hasil penjumlahan tersebut dibagi dengan jumlah data dalam klaster C0. Proses ini bertujuan untuk menemukan pusat klaster yang merupakan representasi rata-rata dari seluruh data dalam klaster tersebut. *Centroid* baru (*C*) dihitung dengan mempertimbangkan jumlah data (*x*) dan banyaknya data (*y*) dalam kolom tertentu (*b*) pada kluster

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

tertentu (*a*). Proses ini membantu dalam menentukan pusat dari setiap kluster berdasarkan distribusi data yang ada. Maka berdasarkan Pers. (2) terbentuk *centroid* baru pada Tabel 6.

$$C_{ab} = \frac{y}{x} \quad (2)$$

Tabel 5 Penentuan *Cluster* dari Nilai Jarak Euclidean Terkecil pada Iterasi-1

Wilayah	2021	2022	2023	*Cluster*
Pacitan	1.067.984	3.420.377	525.663	C2
Ponorogo	424.819	2.777.184	1.168.863	C0
Trenggalek	813.807	3.166.200	779.840	C2
Tulungagung	180.491	2.532.836	1.413.207	C0
Blitar	53.486	2.299.046	1.646.996	C0
Lumajang	125.330	2.477.585	1.468.464	C0
Bondowoso	741.525	3.093.922	852.124	C0
Situbondo	897.670	3.250.067	695.977	C2
Probolinggo	85.943	2.438.298	1.507.756	C0
Mojokerto	131.566	2.483.926	1.462.115	C0
Jombang	219.054	2.133.378	1.812.664	C0
Nganjuk	160.221	2.512.602	1.433.438	C0
Madiun	782.613	3.134.988	811.065	C0
Magetan	920.026	3.272.420	673.621	C2
Ngawi	575.220	2.927.617	1.018.426	C0
Bojonegoro	182.874	2.169.523	1.776.518	C0
Tuban	0	2.352.397	1.593.646	C0
Lamongan	280.896	2.071.749	1.874.335	C0
Gresik	213.436	2.139.076	1.806.976	C0
Bangkalan	213.144	2.565.147	1.380.964	C0
Sampang	390.339	2.742.708	1.203.336	C0
Pamekasan	609.187	2.961.584	984.458	C0
Sumenep	126.277	2.478.667	1.467.374	C0
Kota Kediri	1.593.646	3.946.040	0	C2
Kota Blitar	1.831.710	4.184.102	238.067	C2
Kota Malang	629.484	2.981.875	964.189	C0
Probolinggo	1.673.686	4.026.079	80.048	C2
Kota Pasuruan	1.728.592	4.080.985	134.952	C2
Kota Mojokerto	1.862.217	4.214.611	268.572	C2
Kota Madiun	1.749.898	4.102.290	156.263	C2
Kota Batu	1.719.518	4.071.911	125.881	C2
Kediri	773.341	1.579.061	2.366.979	C0
Malang	2.557.078	204.684	4.150.722	C1
Jember	2.352.397	0	3.946.040	C1
Banyuwangi	904.495	1.447.914	2.498.129	C0
Pasuruan	709.220	1.643.189	2.302.864	C0
Sidoarjo	1.548.154	804.251	3.141.800	C1
Kota Surabaya	1.067.984	3.420.377	525.663	C1

Tabel 6 *Centroid* Baru

Centroid	2021	2022	2023
C0	1.155.111	1.163.356	1.171.476
C1	2.547.718	2.561.061	2.574.058
C2	374.579	377.140	379.657

Langkah 3.4.3 dan 3.4.4 dilakukan terus menerus sampai tidak ada lagi pergeseran nilai jarak dan pusat klaster serta data klaster. Hal ini dilakukan untuk memastikan bahwa klaster yang

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

terbentuk sudah stabil dan tidak berubah lagi. Pergeseran data merujuk pada perubahan posisi data dalam klaster. Pada awalnya, data akan dikelompokkan ke dalam klaster yang memiliki jarak terdekat dengan data tersebut. Namun, setelah pusat klaster dihitung kembali, beberapa data mungkin akan berpindah ke klaster lain yang memiliki jarak lebih dekat. Pergeseran data ini akan terus terjadi sampai tidak ada lagi perubahan posisi data dalam klaster.

Dalam proses 3.4.3, jarak antara data dengan pusat klaster dihitung menggunakan persamaan jarak Euclidean. Pada langkah 3.4.3 terbentuk Tabel 4 yang mengelompokkan data ke dalam klaster yang memiliki jarak terdekat dengan data tersebut. Sedangkan proses 3.4.4 dilakukan untuk menghitung kembali pusat klaster dengan keanggotaan klaster yang baru. Jika pusat klaster tidak berubah, maka proses klaster dianggap selesai. Namun, jika pusat klaster masih berubah, maka proses 3.4.3, dan 3.4.4 akan diulang kembali sampai tidak ada lagi perubahan dalam klaster.

3.4.5 Hasil Akhir Klaster

Tabel 7 Hasil Akhir Klaster

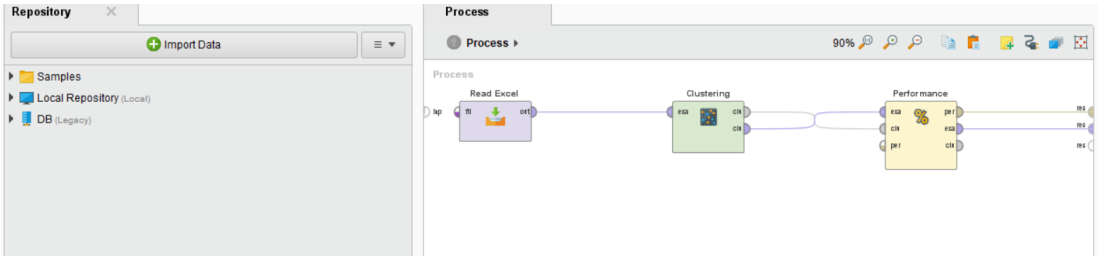
Kota	2021	2022	2023	*Cluster*
Ponorogo	955.839	964.253	972.582	C0
Tulungagung	1.096.588	1.105.337	1.113.973	C0
Blitar	1.231.013	1.240.322	1.249.497	C0
Kediri	1.644.400	1.656.020	1.667.450	C0
Lumajang	1.127.094	1.137.227	1.147.261	C0
Banyuwangi	1.718.462	1.731.731	1.744.814	C0
Probolinggo	1.155.894	1.159.965	1.163.859	C0
Pasuruan	1.611.805	1.619.035	1.626.029	C0
Mojokerto	1.125.522	1.133.584	1.141.516	C0
Jombang	1.325.914	1.335.972	1.345.886	C0
Nganjuk	1.109.683	1.117.033	1.124.247	C0
Ngawi	873.346	877.432	881.393	C0
Bojonegoro	1.307.602	1.315.125	1.322.474	C0
Tuban	1.203.127	1.209.543	1.215.795	C0
Lamongan	1.356.027	1.371.509	1.386.941	C0
Gresik	1.320.570	1.332.664	1.344.648	C0
Bangkalan	1.071.712	1.086.620	1.101.556	C0
Sampang	976.020	984.162	992.210	C0
Pamekasan	853.507	857.818	862.009	C0
Sumenep	1.129.822	1.136.632	1.143.295	C0
Kota Malang	844.933	846.126	847.182	C0
Malang	2.668.296	2.685.900	2.703.175	C1
Jember	2.550.360	2.567.718	2.584.771	C1
Sidoarjo	2.091.930	2.103.401	2.114.588	C1
Kota Surabaya	2.880.284	2.887.223	2.893.698	C1
Pacitan	589.108	592.916	596.649	C2
Trenggalek	734.888	739.669	744.358	C2
Bondowoso	778.525	781.417	784.192	C2
Situbondo	688.337	691.260	694.081	C2
Madiun	750.143	757.665	765.135	C2
Magetan	674.133	678.343	682.466	C2
Kota Kediri	287.962	289.418	290.836	C2
Kota Blitar	150.371	151.960	153.541	C2
Kota Probolinggo	241.202	243.200	245.174	C2
Kota Pasuruan	209.528	211.497	213.450	C2
Kota Mojokerto	133.272	134.350	135.414	C2
Kota Madiun	196.917	199.192	201.460	C2
Kota Batu	214.653	216.735	218.802	C2

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

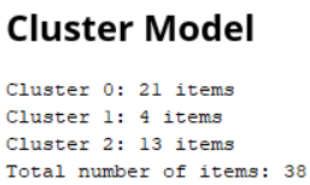
Setelah menyelesaikan proses pada tahap 3.4.3 dan 3.4.4, tidak ditemukan adanya perubahan posisi data dalam klaster. Hal ini menunjukkan bahwa algoritma K-Means telah mencapai konvergensi, di mana posisi setiap data dalam klaster sudah stabil dan tidak lagi berpindah. Oleh karena itu, data yang diperoleh pada Tabel 7 dapat dianggap sebagai hasil akhir dari proses klasterisasi yang telah dilakukan.

3.5 Implementasi K-Means Menggunakan RapidMiner

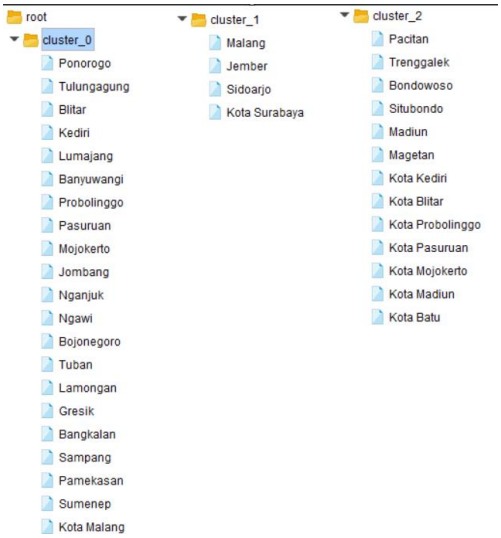
Dalam implementasi algoritma K-Means menggunakan perangkat lunak RapidMiner, *dataset* pada Tabel 1 dijadikan sebagai data yang akan dianalisis yang tersedia dalam format file Excel. Data tersebut kemudian diimpor ke dalam RapidMiner untuk memulai proses analisis. Langkah-langkah analisis yang dilakukan dapat divisualisasikan pada Gambar 3, yang memberikan gambaran jelas dan memudahkan pemahaman tentang proses analisis yang dilakukan.



Gambar 3 Tampilan Input *Dataset* dan Pengujian Menggunakan Performance



Gambar 4 Jumlah Klaster



Gambar 5 Sebaran Klaster

Dari Gambar 3, terlihat bahwa menu performance, yang ditunjukkan oleh kotak berwarna kuning, digunakan untuk membuat grafik metode Elbow dari setiap nilai k yang digunakan. Grafik Elbow *method* ini menunjukkan bagaimana jumlah perbedaan antara titik data dan pusat klaster berubah

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

seiring dengan penambahan jumlah klaster. Tujuan dari grafik ini adalah untuk menemukan titik di mana penurunan perbedaan antara titik data dan pusat klaster mulai melambat secara signifikan, membentuk titik yang menyerupai siku atau "*elbow*", yang menandakan jumlah klaster yang optimal.

Hasil dari perhitungan menggunakan RapidMiner dan perhitungan manual menunjukkan bahwa kedua metode tersebut menghasilkan data yang sama baik dalam jumlah klaster maupun posisi klaster yang dihasilkan. Informasi mengenai jumlah data dalam setiap klaster dan total keseluruhan data dapat ditemukan pada Gambar 4.

Sebaran data pada setiap *cluster* dapat dilihat pada Gambar 5, di mana tergambar dengan jelas sebaran data pada masing-masing *cluster* yang dihasilkan oleh algoritma K-Means. Setiap *cluster* menampung sejumlah nama kota yang sesuai dengan karakteristiknya, mulai dari *cluster* 0, *cluster* 1, hingga *cluster* 2. Hal ini mengindikasikan bahwa algoritma K-Means telah berhasil mengelompokkan kota-kota dalam *dataset* ke dalam tiga kelompok yang berbeda berdasarkan atribut-atribut yang dimiliki.

4. KESIMPULAN

Dalam pengolahan data jumlah penduduk di provinsi jawa timur menggunakan algoritma k-means melalui RapidMiner dan perhitungan manual, ditemukan kesimpulan bahwa kedua metode tersebut menghasilkan jumlah klaster yang serupa yaitu sebanyak 3 klaster dengan *Cluster* 0, yaitu kota dengan sebaran penduduk sedang, mencakup 21 kota, *Cluster* 1, yaitu kota dengan sebaran penduduk terbesar, mencakup 4 kota dan *Cluster* 2, yaitu kota dengan sebaran penduduk terkecil, mencakup 13 kota. Hasil analisis klasterisasi memberikan gambaran yang konsisten terkait pola sebaran penduduk di setiap kota. Kesamaan ini memberikan validitas terhadap aplikasi algoritma k-means dalam konteks analisis demografis provinsi jawa timur.

Oleh karena itu, hasil penelitian ini memberikan keyakinan bahwa algoritma k-means dengan metode Elbow dapat digunakan secara efektif untuk mengklasterisasi data jumlah penduduk, dengan hasil yang konsisten baik melalui pendekatan RapidMiner maupun perhitungan manual. Hasil temuan ini tidak hanya memberikan gambaran yang lebih jelas tentang pola persebaran penduduk di Jawa Timur, tetapi juga berkontribusi memudahkan pemerintah untuk merancang kebijakan yang lebih tepat sasaran berdasarkan pola sebaran dari setiap kota. Ini akan membantu dalam mengalokasi sumber daya yang lebih efisien, perencanaan tata ruang kota yang berkelanjutan, pengelolaan resiko bencana, serta merencanakan pembangunan infrastruktur yang sesuai dengan kebutuhan masing-masing kota sehingga kebijakan-kebijakan yang dilakukan lebih tepat sasaran pada setiap kota di Provinsi Jawa Timur.

DAFTAR PUSTAKA

Anjelita, M., Windarto, A. P., & Hartama, D. (2019). Pemanfaatan Data Mining pada Pengelompokan Provinsi Terhadap Pencemaran Lingkungan Hidup. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *3*(1). https://doi.org/10.30865/komik.v3i1.1675

Anufia, B., & Alhamid, T. (2019). *Instrumen Pengumpulan Data*. OSF. https://doi.org/10.31227/OSF.IO/S3KR6

Badan Pusat Statistik. (2024, March 14). *Jumlah Penduduk Provinsi Jawa Timur (Jiwa), 2021-2023*. Badan Pusat Statistik Kota Kediri. https://kedirikota.bps.go.id/indicator/12/358/1/jumlah-penduduk-provinsi-jawa-timur.html

Budiana, N. D., Siregar, R. R. A., & Susanti, M. N. I. (2019). Penetapan Instruktur Diklat Menggunakan Metode Clustering K-Means dan Topsis Pada PT PLN (Persero) Udiklat Jakarta. *PETIR*, *12*(2), 111–121. https://doi.org/10.33322/petir.v12i2.454

Dewi, F. P., Aryni, P. S., & Umaidah, Y. (2022). Implementasi Algoritma K-Means Clustering Seleksi Siswa Berprestasi Berdasarkan Keaktifan dalam Proses Pembelajaran. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *7*(2), 111–121. https://doi.org/10.14421/jiska.2022.7.2.111-121

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Fahrozi, A. Al, Insani, F., Budianita, E., & Afrianty, I. (2023). Implementasi Algoritma K-Means dalam Menentukan Clustering pada Penilaian Kepuasan Pelanggan di Badan Pelatihan Kesehatan Pekanbaru. *Indonesian Journal of Innovation Multidisipliner Research*, *1*(4), 474–492. https://doi.org/10.31004/IJIM.V1I4.53

Fitriyah, H., Safitri, E. M., Muna, N., Khasanah, M., Aprilia, D. A., & Nurdiansyah, D. (2023). Implementasi Algoritma Clustering dengan Modifikasi Metode Elbow untuk Mendukung Strategi Pemerataan Bantuan Sosial di Kabupaten Bojonegoro. *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, *4*(3), 1598–1607. https://doi.org/10.46306/lb.v4i3.453

Harahap, L. M., Fuadi, W., Rosnita, L., Darnila, E., & Meiyanti, R. (2022). Klastering Sayuran Unggulan Menggunakan Algoritma K-Means. *Jurnal Teknik Informatika Dan Sistem Informasi*, *8*(3), 567–579. https://doi.org/10.28932/jutisi.v8i3.5277

Herviany, M., Putri Delima, S., Nurhidayah, T., & Kasini, K. (2021). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokkan Daerah Rawan Tanah Longsor Pada Provinsi Jawa Barat. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *1*(1), 34–40. https://doi.org/10.57152/malcom.v1i1.60

Naldy, E. T., & Andri, A. (2021). Penerapan Data Mining Untuk Analisis Daftar Pembelian Konsumen Dengan Menggunakan Algoritma Apriori Pada Transaksi Penjualan Toko Bangunan MDN. *Jurnal Nasional Ilmu Komputer*, *2*(2), 89–101. https://doi.org/10.47747/jurnalnik.v2i2.525

Nawassyarif, M. Julkarnain, & Rizki Ananda, K. (2020). Sistem Informasi Pengolahan Data Ternak Unit Pelaksana Teknis Produksi dan Kesehatan Hewan Berbasis Web. *Jurnal Informatika, Teknologi Dan Sains*, *2*(1), 32–39. https://doi.org/10.51401/jinteks.v2i1.556

Nofiar, A., Defit, S., & Sumijan. (2019). Penentuan Mutu Kelapa Sawit Menggunakan Metode K-Means Clustering. *Jurnal KomtekInfo*, *5*(3), 1–9. https://doi.org/10.35134/komtekinfo.v5i3.26

Sadewo, M. G., Windarto, A. P., & Wanto, A. (2018). Penerapan Algoritma Clustering dalam Mengelompokkan Banyaknya Desa/Kelurahan Menurut Upaya Antisipasi/Mitigasi Bencana Alam Menurut Provinsi dengan K-Means. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *2*(1), 311–319. https://doi.org/10.30865/komik.v2i1.943

Sholeh, M., Suraya, S., & Andayati, D. (2022). Machine Linear untuk Analisis Regresi Linier Biaya Asuransi Kesehatan dengan Menggunakan Python Jupyter Notebook. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *8*(1), 20–27. https://doi.org/10.26418/JP.V8I1.48822

Syahfitri, N., Budianita, E., Nazir, A., & Afrianty, I. (2023). Pengelompokan Produk Berdasarkan Data Persediaan Barang Menggunakan Metode Elbow dan K-Medoid. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *4*(3), 1668–1675. https://doi.org/10.30865/KLIK.V4I3.1525

Talakua, M. W., Leleury, Z. A., & Taluta, A. W. (2017). Analisis Cluster dengan Menggunakan Metode K-Means untuk Pengelompokkan Kabupaten/Kota di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *11*(2), 119–128. https://doi.org/10.30598/barekengvol11iss2pp119-128

Triandini, M., Defit, S., & Nurcahyo, G. W. (2021). Data Mining dalam Mengukur Tingkat Keaktifan Siswa dalam Mengikuti Proses Belajar pada SMP IT Andalas Cendekia. *Jurnal Informasi Dan Teknologi*, 167–173. https://doi.org/10.37034/jidt.v3i3.120

Triyansyah, D., & Fitrianah, D. (2018). Analisis Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing. *Jurnal Telekomunikasi Dan Komputer*, *8*(3), 163–182. https://doi.org/10.22441/incomtech.v8i3.4174

Virgo, I., Defit, S., & Yuhandri, Y. (2020). Klasterisasi Tingkat Kehadiran Dosen Menggunakan Algoritma K-Means Clustering. *Jurnal Sistim Informasi Dan Teknologi*, *2*(1), 23–28. https://doi.org/10.37034/jsisfotek.v2i1.17

Wicaksana, R. S., Heksaputra, D., Syah, T. A., & Nur'aini, F. F. (2023). Pendekatan K-Means Clustering Metode Elbow Pada Analisis Motivasi Pengunjung Festival Halal JHF#2. *Jurnal Ilmiah Ekonomi Islam*, *9*(3), 4162. https://doi.org/10.29040/jiei.v9i3.10591

Yudhistira, A., & Andika, R. (2023). Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering. *Journal of Artificial Intelligence and Technology Information (JAITI)*, *1*(1), 20–28. https://doi.org/10.58602/jaiti.v1i1.22

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024, Pp. 147 – 158
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Deteksi Pelanggaran pada *Zebra Cross* dengan *Water Spray* dan *Buzzer* berbasis IoT

Dina Uzlifatul Firdaus [(1)*], Febrian Wahyu Christanto [(2)]
Teknik Informatika, Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang, Semarang
e-mail : dina.uzlifatul5@gmail.com, febrian.wahyu.christanto@usm.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 21 FebruarI 2024, direvisi 5 Mei 2024, diterima 7 Mei 2024, dan dipublikasikan 25 Mei 2024.

Abstract

A zebra crossing is a road marking indicating a crossing path for pedestrians. Zebra crossings are directly used to signal drivers to stop at the line boundaries. Because the zebra crossing functions as a crossing area, pedestrians and motorized vehicle drivers must understand and obey existing traffic signs. According to data from the WHO (World Health Organization), 270,000 pedestrians die every year or around 22% of all victims who die due to road accidents. An ESP32-Cam microcontroller, an E18-D80NK Infrared Proximity Sensor, water spray and buzzer approaches, and the prototype development method were used to design a system for detecting crossing violations at zebra crossings to address this issue. The Infrared Proximity sensor will automatically detect when a crossing violation occurs, then the water spray will spray water, and the buzzer will make a sound as a warning sign to obey traffic. ESP32-Cam functions as an image capturer if a crossing violation has occurred and is automatically sent to the Telegram Bot. The confusion matrix test tested the research results with an accuracy value of 83.33%, a precision value of 83.33%, and a recall value of 88.23%.

Keywords: Buzzer, ESP32-Cam, E18-D80NK Infrared Proximity Sensor, Water Spray, Zebra Crossing

Abstrak

Zebra cross adalah marka jalan yang menandakan jalur penyeberangan bagi pejalan kaki untuk melintas. Secara langsung *zebra cross* juga digunakan untuk penanda pengendara untuk berhenti pada batasan garisnya. Karena fungsi *zebra cross* sebagai area penyeberangan, maka baik pejalan kaki ataupun pengendara kendaraan bermotor wajib memahami dan mematuhi rambu-rambu lalu lintas yang ada. Menurut data dari WHO (*World Health Organization*) terdapat 270.000 pejalan kaki meninggal dunia setiap tahun atau sekitar 22% dari seluruh korban meninggal akibat kecelakaan di jalan. Untuk menangani permasalahan tersebut maka dibuatlah sebuah sistem pendeteksi pelanggaran penyeberangan pada *zebra cross* dengan metode pengembangan *Prototype* serta mikrokontroler ESP32-Cam dan Sensor *Infrared Proximity* E18-D80NK menggunakan teknik *water spray* dan *buzzer*. Sensor *Infrared Proximity* akan mendeteksi secara otomatis saat pelanggaran penyeberangan terjadi dan kemudian *water spray* akan menyemprotkan air serta *buzzer* akan mengeluarkan suara sebagai tanda peringatan untuk mematuhi lalu lintas. ESP32-Cam berfungsi sebagai penangkap gambar jika telah terjadi pelanggaran penyeberangan dan secara otomatis dikirim ke *Bot* Telegram. Hasil penelitian diuji menggunakan *confusion matrix* dan didapatkan nilai *accuracy* sebesar 83,33%, *precision* sebesar 83,33%, dan *recall* sebesar 88,23%.

Kata Kunci: *Buzzer*, ESP32-Cam, Sensor *Infrared Proximity* E18-D80NK, *Water Spray*, *Zebra Cross*

1. PENDAHULUAN

Trotoar dan *zebra cross* merupakan salah satu bentuk fasilitas bagi pejalan kaki yang sangat diperlukan untuk menjamin keselamatan dan kenyamanan masyarakat dalam melakukan aktifitasnya sehari-hari sebagai alternatif menuju tempat tujuan (Nugroho, 2018). Tampilan visual *zebra cross* diwakili oleh marka lajur berupa garis vertikal hitam putih. *Zebra cross* adalah hak

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

jalan bagi pejalan kaki dan tidak boleh dilanggar oleh pengemudi lain yang melintas. *Zebra cross* digunakan sebagai sinyal bagi pengemudi agar memperlambat kecepatan pada saat pejalan kaki menyeberang. Setiap pengemudi kendaraan tidak diperkenankan berhenti pada garis *zebra cross* yang memang diperuntukkan untuk pejalan kaki (Rahmawati et al., 2022). Peraturan mengenai tidak berhenti di *zebra cross* sudah diatur dalam Undang-Undang Lalu Lintas dan Angkutan Jalan (UULAJ) Nomor 22 Tahun 2019 pasal 284 yang berbunyi "Setiap orang yang mengemudikan kendaraan bermotor dengan tidak mengutamakan keselamatan pejalan kaki atau pesepeda sebagaimana yang dimaksud dalam pasal 106 ayat (2) dipidana kurungan paling lama 2 (dua) bulan dan denda sebesar Rp 500.000 (lima ratus ribu rupiah)" (Undang-Undang Republik Indonesia Nomor 22 Tahun 2009 Tentang Lalu Lintas Dan Angkutan Jalan, 2009) (Alamsyah, 2019).

Namun masih banyak kendaraan yang melanggar peraturan tersebut sehingga menyulitkan pejalan kaki untuk menyeberang jalan karena kendaraan tersebut menghalangi jalan (Rahmawati et al., 2022). Faktor utama yang perlu diperhatikan dalam penyediaan sarana penyeberangan bagi pejalan kaki adalah faktor keselamatan. Sarana tersebut harus mampu menjamin keselamatan pejalan kaki di jalan. Namun pada kenyataannya masih banyak pejalan kaki yang menjadi korban kecelakaan lalu lintas, dikarenakan ketidakpatuhan maupun ketidakdisiplinan pengguna jalan. Peluang dan frekuensi terjadinya kecelakaan dapat diamati dari dua sisi, yaitu dari sisi pengemudi kendaraan bermotor dan pejalan kaki itu sendiri (Widyaningsih & Daniel, 2019).

Kewajiban pejalan kaki untuk menyeberang jalan pada tempat yang telah ditentukan diatur berdasarkan pada pasal 132 Undang Undang Nomor 22 Tahun 2009 (2009) Tentang Lalu Lintas dan Angkutan Jalan yang menjelaskan bahwa "Pejalan kaki wajib menggunakan bagian jalan yang diperuntukan bagi pejalan kaki atau jalan yang paling tepi dan menyeberang jalan di tempat yang telah ditentukan." Adanya peraturan ini jelas menunjukkan bahwa apabila sudah disediakan atau ditentukan di mana tempat seharusnya khusus bagi pejalan kaki untuk menyeberang jalan. Pejalan kaki yang menyeberang jalan tidak pada tempatnya harus dihukum karena peraturan sudah tertulis dengan jelas dan fasilitas telah disediakan (Embrianto & Sulistyowati, 2020). Namun, seringkali orang mengalami kecelakaan karena faktor dari pejalan kaki tersebut. Menurut data dari WHO (World Health Organization) terdapat 270.000 pejalan kaki meninggal dunia setiap tahunnya atau setara dengan 22% dari seluruh korban meninggal akibat kecelakaan lalu lintas. Sedangkan di Indonesia sendiri, mengutip dari laman Global Road Safety Facility, persentase kematian pejalan kaki akibat kecelakaan lalu lintas sebesar 38% dari total 31.282 kematian di jalan raya yang dilaporkan pada tahun 2016 (Wicaksono et al., 2021).

Kecelakaan tersebut disebabkan karena pejalan kaki yang menyeberang jalan tidak memperhatikan rambu lalu lintas, biasanya karena pandangan mereka teralihkan, misalnya karena terlalu fokus pada ponsel sehingga tidak memperhatikan keadaan sisi kiri dan kanannya sebelum mereka menginjakkan kaki di *zebra cross* tersebut (Yolanda et al., 2021). Berdasarkan permasalahan tersebut, salah satu hal yang dapat dimanfaatkan untuk mengurangi jumlah pelanggaran pengguna jalan pada *zebra cross* yaitu dengan sistem pendeteksi penyeberangan pada *zebra cross*. Penelitian ini dibuat menggunakan ESP32-Cam dan Sensor Infrared Proximity E18-D80NK dengan teknik *water spray* dan *buzzer*. Sensor Infrared Proximity E18-D80NK digunakan untuk mendeteksi objek dalam jarak 3-80 cm secara otomatis pada saat lampu lalu lintas dalam keadaan berwarna hijau, kemudian *water spray* berfungsi sebagai penyemprotan air gunanya untuk memberikan efek visual kepada pejalan kaki dan *buzzer* akan mengeluarkan suara alarm sebagai tanda peringatan untuk pejalan kaki agar mematuhi lalu lintas. ESP32-Cam berfungsi sebagai penangkap gambar jika telah terjadi pelanggaran penyeberangan dan akan secara otomatis dikirim ke aplikasi pesan Telegram melalui Bot Telegram. Selanjutnya, Sensor Infrared Proximity E18-D80NK akan mati secara otomatis jika lampu lalu lintas dalam keadaan berwarna merah dan kuning. Sistem deteksi ini saling berhubungan dengan *traffic light*, dan akan aktif apabila *traffic light* dalam keadaan menyala dan berwarna hijau. Tujuan dari penelitian ini adalah untuk merancang sebuah sistem pendeteksi pelanggaran penyeberangan yang dibuat untuk pejalan kaki yang akan melewati *zebra cross* dengan bekerja secara otomatis dan

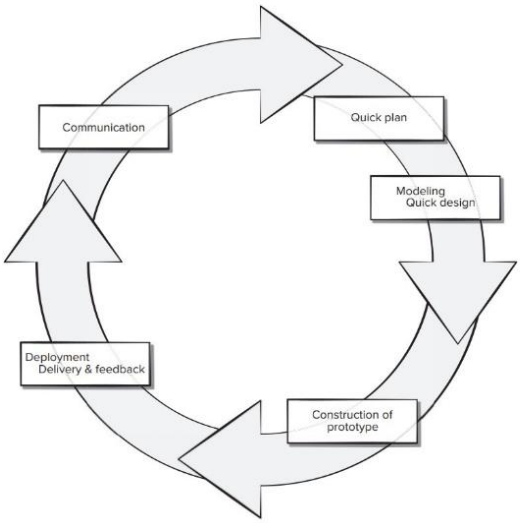
Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

memberikan peringatan secara terus menerus kepada pejalan kaki agar tidak melakukan pelanggaran rambu lalu lintas serta mematuhi peraturan yang ada.

2. METODE PENELITIAN

2.1 Metode Pengembangan Perangkat

Penelitian ini menggunakan metode pengembangan Prototype yang terdiri dari 5 (lima) tahap yaitu *Communication*, *Quick Plan*, *Modelling Quick Design*, *Construction of Prototype*, dan *Deployment Delivery & Feedback* (Pressman & Maxim, 2019). Gambar 1 merupakan metode pengembangan *Prototype* yang digunakan pada penelitian ini. Langkah pertama tahap yang dilakukan adalah *Communication*, pada tahap ini dilakukan pemikiran terhadap situasi yang ada di sekitar lampu lalu lintas. Seperti halnya pejalan kaki menyeberang pada saat lampu lalu lintas dalam keadaan berwarna hijau dan terkadang pejalan kaki menyeberang dengan pandangan tidak fokus atau pandangan teralihkan oleh ponsel. Tahapan kedua adalah *Quick Plan*, dari permasalahan yang ada ditentukan rencana bagaimana cara merancang sistem pendeteksi pelanggaran penyeberangan pada *zebra cross* yang mampu mendeteksi secara otomatis dengan teknik *water spray* dan *buzzer*, serta bagaimana cara memberikan peringatan kepada pejalan kaki agar tidak melakukan pelanggaran rambu lalu lintas.



Gambar 1 Metode Pengembangan Prototype

Tahapan ketiga adalah *Modeling Quick Design*, pembuatan rancangan alat-alat yang akan dibuat seperti Sensor Infrared Proximity E18-D80NK yang digunakan untuk mendeteksi objek, *buzzer* digunakan sebagai alarm, modul *traffic light* sebagai acuan sensor aktif dan tidaknya, ESP32-Cam sebagai mikrokontroler, dan *water spray pump* digunakan untuk menyemprotkan air sebagai tanda telah terjadinya pelanggaran. Tahapan keempat adalah *Construction of Prototype*, sistem yang telah selesai dibuat, kemudian akan dilakukan pengujian, apakah sistem tersebut layak digunakan atau tidak. Seperti pengujian deteksi pada sensor infrared terhadap objek, untuk memastikan bahwa sensor telah bekerja. Kemudian pengujian terhadap lampu lalu lintas (*traffic light*), untuk memastikan jika lampu dalam keadaan berwarna hijau maka sensor akan menyala, sedangkan jika lampu dalam keadaan berwarna merah dan kuning, maka sensor akan mati. Dan pengujian terhadap *buzzer* dan *water spray*, untuk memastikan bahwa *buzzer* akan mengeluarkan bunyi alarm dan *water spray* akan menyemprotkan air secara otomatis jika objek melewati sensor yang aktif.

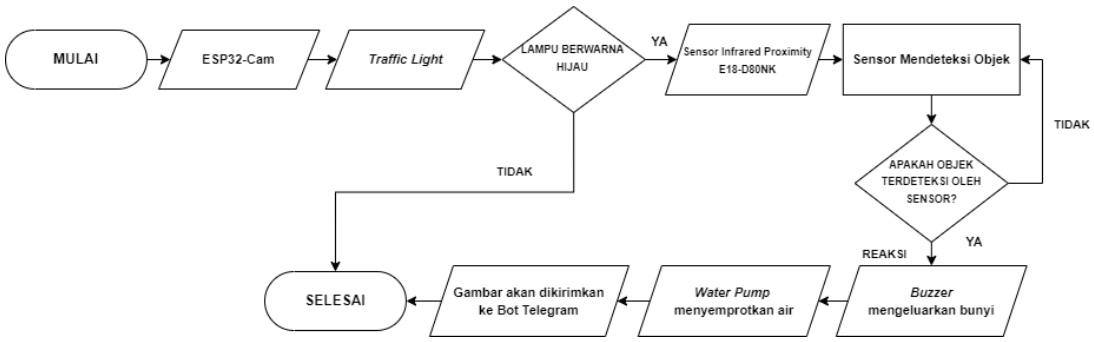
Tahapan kelima adalah *Deployment Delivery & Feedback*, tahap ini merupakan tahap akhir dari pengembangan sistem deteksi pelanggaran penyeberangan pada *zebra cross*. Proses evaluasi bisa didapat pada saat setelah melakukan pengujian oleh penulis. Apabila sistem sudah

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

mendapatkan kelayakan pada saat pengujian khususnya penentuan presentasi keakuratan untuk melihat kerja Sensor Infrared Proximity E18-D80NK, *water spray*, dan *buzzer* menyala atau tidak.

2.2 Alur Kerja Penelitian

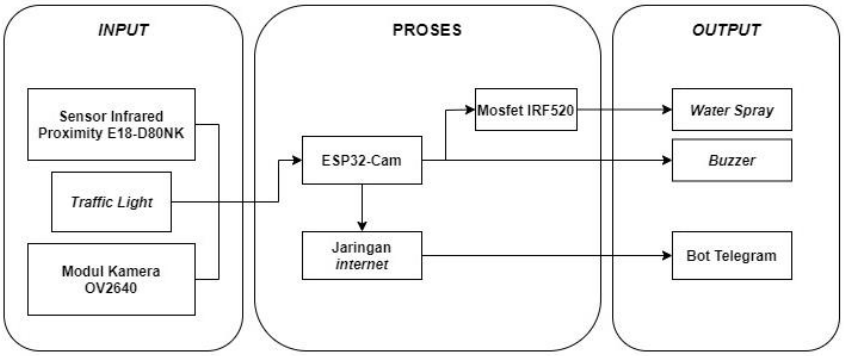
Penelitian ini disusun dengan beberapa langkah yang akan dilakukan secara sistematis. Gambar 2 menjelaskan alur penelitian dari sistem deteksi pelanggaran penyeberangan pada *zebra cross*. ESP32-Cam merupakan proses yang pertama kali saat alat dinyalakan dan terhubung dengan jaringan internet, *traffic light* sebagai acuan sensor aktif dan tidaknya, jika lampu dalam keadaan berwarna hijau maka Sensor Infrared proximity E18-D80NK akan aktif mendeteksi objek pejalan kaki, ketika sensor berhasil mendeteksi maka *water pump* akan menyemprotkan air dan *buzzer* akan mengeluarkan bunyi sebagai tanda pelanggaran penyeberangan telah terjadi. Kemudian hasil deteksi dari sensor yang aktif, maka ESP32-Cam akan menangkap gambar dan dikirimkan ke aplikasi Telegram melalui Bot Telegram.



Gambar 2 Alur Penelitian

2.3 Diagram Blok

Dalam penelitian ini diagram blok menjelaskan proses kerja alat atau sistem sebagai rancangan awal sebelum dibuat. Dengan adanya diagram blok maka semuanya akan terlihat jelas seperti komponen yang digunakan seperti *input*, proses, dan *output* sistem. Adapun diagram blok dari sistem deteksi pelanggaran penyeberangan pada *zebra cross* terdapat pada Gambar 3.



Gambar 3 Diagram Blok Sistem

Diagram blok sistem dibagi menjadi 3 (tiga) bagian yaitu *input*, proses, dan *output*. *Input* pada sistem adalah Sensor Infrared Proximity E18-D80NK yang berfungsi untuk mendeteksi suatu objek berupa pejalan kaki yang hendak menyeberang jalan, *Traffic light* sebagai pengontrolan jalannya lalu lintas, dan Modul Kamera OV2640 akan mengidentifikasi objek yang kemudian dilanjutkan mikrokontroler ESP32-Cam. Pada proses, ESP32-Cam akan mengolah data yang diterima dari Sensor Infrared Proximity dengan koneksi dari jaringan internet. Mosfet IRF520

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

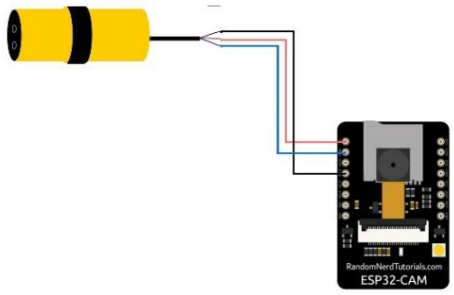
sebagai penggerak pada *water pump* dan *buzzer*. Pada *output*, *water pump* akan menyemprotkan air secara otomatis dan *buzzer* akan mengeluarkan suara jika terdeteksi pelanggaran penyeberangan pada *zebra cross*. Kemudian, tangkapan gambar yang telah diterima ESP32-Cam akan dikirimkan ke aplikasi Telegram melalui Bot Telegram.

2.4 Perancangan Perangkat Keras

Tahap perancangan perangkat keras (*hardware*) merupakan tindak lanjut dari Analisa. Tahapan ini menghasilkan suatu perancangan sistem yang diperlukan dalam pendeteksian pelanggaran penyeberangan pada *zebra cross*. Beberapa rangkaian sistem yang digunakan pada penelitian ini dijabarkan dalam berbagai skema berikut.

2.4.1 Skema Rangkaian ESP32-Cam dengan Sensor Infrared Proximity E18-D80NK

Modul ESP32-Cam memiliki dua sisi dalam rangkaian modulnya. Di bagian atas terdapat modul kamera yang dapat dibongkar pasang dan ada microSD yang dapat diisi. Di bagian belakang modul terdapat antena internal, konektor untuk antena eksternal, pin *male* untuk I/O dan ESP32 sebagai otaknya (Mahmuddin et al., 2023). Sensor Infrared Proximity E18-D80NK memiliki tiga kabel dengan rincian kabel biru dihubungkan ke *ground*, kabel coklat dihubungkan ke tegangan 5V, dan kabel hitam dihubungkan ke *out* (Wahyudi & Aziz, 2022). Gambar 4 berikut merupakan skema rangkaian ESP32-Cam dengan Sensor Infrared Proximity E18-D80NK.



Gambar 4 Skema ESP32-Cam dengan Sensor Infrared Proximity E18-D80NK

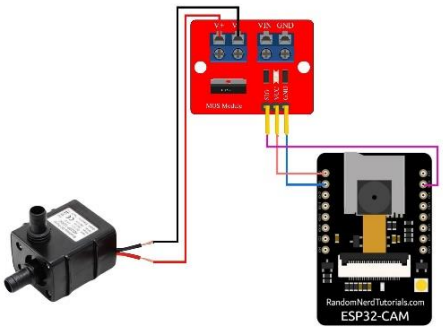
Gambar 4 menjelaskan skema rangkaian dari ESP32-Cam dengan Sensor Infrared Proximity E18-D80NK, konfigurasi ESP32-Cam pin GND dengan Sensor Infrared Proximity E18-D80NK pin GND, konfigurasi ESP32-Cam pin 5V dengan Sensor Infrared Proximity E18-D80NK pin VCC, dan konfigurasi ESP32-Cam pin GPIO 13 dengan Sensor Infrared Proximity E18-D80NK pin *out*.

2.4.2 Skema Rangkaian ESP32-Cam dengan Mosfet IRF520 dan *Water pump*

ESP32-Cam ini memiliki berbagai pin GPIO (*General Purpose Input/Output*) yang dapat diprogram sesuai kebutuhan, pin ini dapat digunakan untuk mengendalikan dan berkomunikasi dengan berbagai perangkat eksternal melalui protokol I2C, SPI, UART, dan PWM (IOTkece, 2021). Mosfet IRF520 ini merupakan modul untuk mempermudah penggunaan transistor Mosfet IRF520 yang *driver* ini memiliki switching time yang tinggi, artinya perubahan dari *low* ke *high* dan sebaliknya sangat cepat, sehingga cocok untuk kontrol *switching* tegangan yang lebih tinggi (Al Khaledi et al., 2022). *Water pump* ini menggunakan motor DC *brushless* dan bekerja dengan tegangan DC 12V 240L/jam, kelebihan dari *water pump* ini adalah tidak berisik saat digunakan dan aman saat bekerja di air (Ulum et al., 2022).

ESP32-Cam pada Gambar 5 menjelaskan skema rangkaian dari ESP32-Cam dengan Mosfet IRF520 dan *Water pump*, konfigurasi Mosfet IRF520 pin VCC dengan ESP32-Cam pin 5V, konfigurasi Mosfet IRF520 pin GND dengan ESP32-Cam pin GND, konfigurasi Mosfet IRF520 pin SIG dengan ESP32-Cam pin GPIO 16, konfigurasi Mosfet IRF520 pin V+ dengan kabel merah *water pump*, dan konfigurasi Mosfet IRF520 pin V- dengan kabel hitam *water pump*.

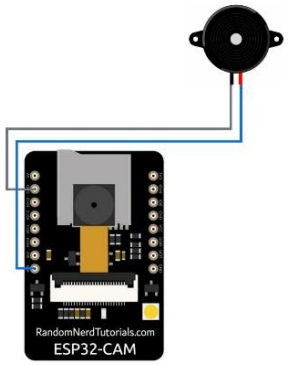
Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.



Gambar 5 Skema ESP32-Cam dengan Mosfet IRF520 dan *Water Pump*

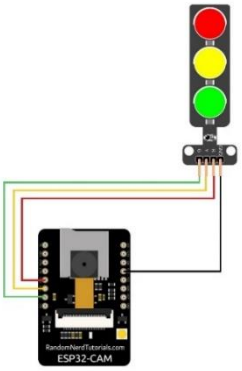
2.4.3 Skema Rangkaian ESP32-Cam dengan *Buzzer*

Buzzer adalah sebuah komponen yang memiliki fungsi mengubah arus listrik menjadi suara. Dan pada dasarnya prinsip kerja *buzzer* hampir sama dengan *speaker*. *Buzzer* terdiri dari sebuah diafragma yang memiliki kumparan. Karena kumparan dipasang pada diafragma maka setiap getaran diafragma secara bolak-balik sehingga membuat udara bergetar dan menghasilkan suara (Prabowo et al., 2020). *Buzzer* biasa digunakan sebagai indikator bahwa proses telah selesai atau terjadi suatu kesalahan pada sebuah alat (alarm) (Hasanah et al., 2021). Gambar 6 menjelaskan skema rangkaian dari ESP32-Cam dengan *buzzer*, konfigurasi ESP32-Cam pin GPIO 4 dengan *buzzer* kaki positif dan konfigurasi ESP32-Cam pin GND dengan *buzzer* kaki negatif.



Gambar 6 Skema ESP32-Cam dengan *Buzzer*

2.4.4 Skema Rangkaian ESP32-Cam dengan *Traffic Light*



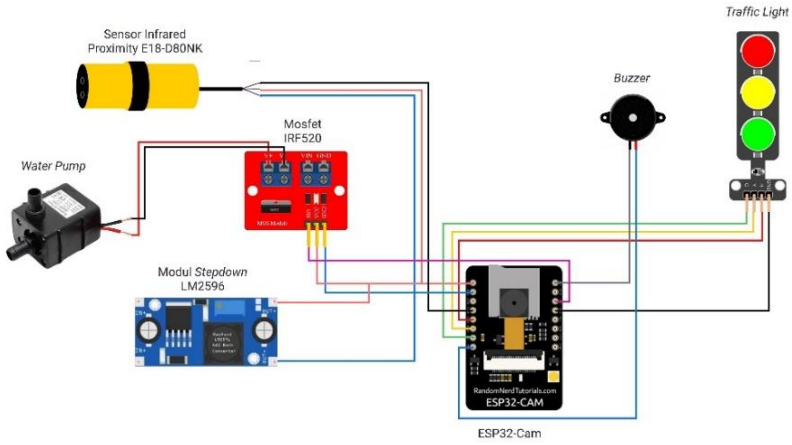
Gambar 7 Skema ESP32-Cam dengan *Traffic Light*

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

ESP32-Cam adalah mikrokontroler berfitur lengkap yang juga memiliki kamera video terintegrasi dan soket kartu microSD. Ini murah dan mudah digunakan untuk perangkat *Internet of Things* (IoT) yang membutuhkan kamera dengan fungsi-fungsi canggih seperti pelacakan dan pengenalan gambar (Saputra & Darujati, 2020). Fungsi *traffic light* digunakan untuk mengatur perpindahan kendaraan di persimpangan jalan agar tidak terjadi kemacetan (Damayanti et al., 2019). Gambar 7 menjelaskan skema rangkaian dari ESP32-Cam dengan *traffic light*, konfigurasi ESP32-Cam pin GPIO 15 dengan *traffic light* pin Red, konfigurasi ESP32-Cam pin GPIO 14 dengan *traffic light* pin Yellow, konfigurasi ESP32-Cam pin GPIO 2 dengan *traffic light* pin Green, dan konfigurasi ESP32-Cam pin GND dengan *traffic light* pin GND.

2.5 Skema Keseluruhan Rangkaian

Skema ini digunakan untuk mengetahui dan menerangkan keseluruhan model dari penelitian ini. Gambar 8 merupakan komponen-komponen yang telah disusun dan dirangkai untuk memberntuk suatu alat atau skema keseluruhan rangkaian dari sistem deteksi pelanggaran penyeberangan pada *zebra cross* dengan teknik *water spray* dan *buzzer* menggunakan ESP32-Cam dan Sensor Infrared Proximity E18-D80NK yang terdapat dalam Gambar 8.



Gambar 8 Skema Keseluruhan Rangkaian

Berdasarkan Gambar 8 merupakan skema rangkaian seperti ESP32-Cam digunakan sebagai penangkap jaringan internet dan penangkap gambar pada saat ada objek atau pejalan kaki yang melanggar, Sensor Infrared Proximity E18-D80NK sebagai sensor untuk mendeteksi objek yaitu pejalan kaki yang akan menyeberang jalan pada *zebra cross*, *buzzer* digunakan untuk mengeluarkan bunyi atau alarm apabila terdeteksi terjadinya pelanggaran, *Water pump* digunakan untuk menyemprotkan air pada saat terdeteksinya pelanggaran, *traffic light* digunakan sebagai lampu lalu lintas. Pada saat lampu dalam keadaan berwarna hijau maka Sensor Infrared Proximity E18-D80NK, *water spray*, dan *buzzer* akan menyala. Kemudian apabila lampu dalam keadaan berwarna merah dan kuning maka Sensor Infrared Proximity E18-D80NK akan mati secara otomatis. Mosfet IRF520 digunakan untuk mengontrol *switching* tegangan dan untuk mendorong beban DC. Modul Stepdown LM2596 digunakan untuk level tegangan arus searah (DC) di bawah tegangan *input*.

2.6 Pengujian

Pada tahap ini dilakukan pengujian sistem menggunakan pengujian *black box* pada *hardware* dengan cara menguji setiap proses yang terjadi pada sistem dan untuk mengetahui keberhasilan sistem berjalan sesuai dengan yang diinginkan. Kemudian untuk mengetahui *delay* pada sistem, dilakukan pengujian *delay* menggunakan *stopwatch*. Sedangkan untuk mengetahui nilai *accuracy*, *precision*, dan *recall* dilakukan pengujian *Confusion matrix*. *Accuracy* menggambarkan seberapa akurat data mengklasifikasikan dengan benar. *Precision* menggambarkan tingkat

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

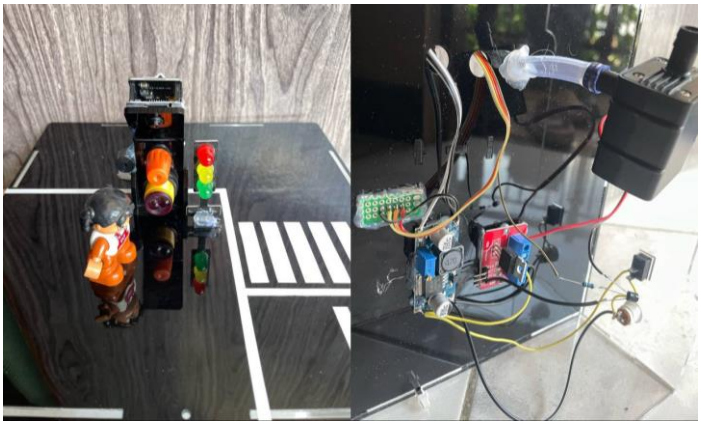
keakuratan antara data yang diinginkan dengan hasil prediksi yang diberikan. Dan *recall* menggambarkan keberhasilan data dalam menemukan kembali sebuah informasi (Düntsch & Gediga, 2019). Dalam *Confusion matrix* digunakan data *testing* sebanyak 30 kali percobaan. *Confusion matrix* dihitung dengan cara melakukan kalkulasi nilai jumlah dari *True Positive*, *True Negative*, *False Positive*, dan *False Negative*. Dengan begitu dapat diperoleh nilai *accuracy*, *precision*, dan *recall* tersebut.

2.7 Evaluasi

Tahapan terakhir dalam penelitian ini yaitu evaluasi. Hasil pengujian yang telah didapatkan akan dihitung untuk diketahui bagaimana kemampuan sistem dalam melakukan pendeteksian pelanggaran penyeberangan pada *zebra cross*. Perhitungan evaluasi mengggunakan *confusion matrix* untuk menghitung nilai *accuracy*, *precision*, dan *recall*.

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini perangkat keras (*hardware*) berfungsi untuk memasukkan data ke *processor* atau untuk menyimpan dan menghasilkan data. Bagian dari perangkat keras (*hardware*) harus saling terhubung agar perintah yang diberikan oleh *processor* dapat berjalan sesuai dengan fungsinya. Gambar 9 merupakan *prototype* rangkaian perangkat keras (*hardware*).



Gambar 9 Prototype Rangkaian Perangkat Keras

Gambar 9 merupakan perakitan rangkaian perangkat keras (*hardware*). Pada pembentukan rangkaian ini di bagian atas terdapat komponen berupa ESP32-Cam, *traffic light*, Sensor Infrared Proximity E18-D80NK, *buzzer*, *water spray*, dan mainan pejalan kaki sebagai objek simulasi. Kemudian pada bagian bawah terdapat komponen berupa pompa air, stepdown LM2596, dan mosfet IRF520.

Pengujian penelitian ini dilakukan dengan pengujian *black box* terhadap komponen dari sistem deteksi pelanggaran penyeberangan pada *zebra cross*. Tabel 1 merupakan hasil dari pengujian *black box* alat seperti Sensor Infrared Proximity E18-D80NK aktif saat mendeteksi objek yang melanggar penyeberangan, *traffic light* sebagai pengontrol jalannya lalu lintas, ESP32-Cam sebagai mikrokontroler sekaligus penangkap gambar, *water spray* dan *buzzer* dapat bekerja dengan baik pada saat terdeteksinya pelanggaran penyeberangan pada *zebra cross* terjadi. Dengan begitu hasil pengujian menggunakan *black box* ini sistem dapat berjalan baik dan tidak ditemukan masalah.

Pada penelitian ini dilakukan pula pengujian menggunakan *stopwatch* terhadap komponen *output* seperti *water spray* dan *buzzer* sebagai pendeteksi pelanggaran penyeberangan pada *zebra cross* pada cuaca cerah. Tabel 2 adalah pengujian *delay* respon *water spray* dan *buzzer*. Berdasarkan Tabel 2 dilakukan pengujian *delay* respon menggunakan *stopwatch* dan dengan percobaan sebanyak 10 kali pelanggaran. Dari hasil perhitungan tersebut, didapatkan nilai

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

terendah pada *delay* respon *water spray* 0,87 detik, dan nilai terendah pada *delay* respon *buzzer* 0,84 detik. Kemudian didapatkan nilai tertinggi pada *delay* respon *water spray* 1,01 detik, dan nilai tertinggi pada *delay* respon *buzzer* 0,93 detik. Lalu didapatkan untuk rata-rata *delay* keseluruhan pada saat terjadi pelanggaran penyeberangan pada *zebra cross* adalah 0,85 detik.

Tabel 1 Pengujian *Black Box*

No.	Pengujian	Pengamatan	Hasil	Keterangan
1	Sensor Infrared Proximity E18-D80NK	Sensor Infrared Proximity E18-D80NK dapat mendeteksi objek yaitu pejalan kaki yang melanggar penyeberangan pada saat lampu lalu lintas dalam keadaan berwarna hijau.	Berhasil	Ketika objek yaitu pejalan kaki melewati Sensor Infrared Proximity E18-D80NK, maka LED pada Sensor Infrared Proximity E18-D80NK akan menyala dan Sensor Infrared Proximity E18-D80NK berada pada kondisi *HIGH*. Jika tidak ada yang melewati Sensor Infrared Proximity E18-D80NK, maka LED pada Sensor Infrared Proximity E18-D80NK mati.
2	*Traffic Light*	LED merah menyala selama 5 detik, LED kuning menyala selama 1 detik, dan LED hijau menyala selama 10 detik.	Berhasil	*Traffic light* sebagai pengontrolan jalannya lalu lintas. LED merah dan kuning menyala sebagai tanda Sensor Infrared Proximity E18-D80NK mati, maka objek pejalan kaki dapat menyeberang. Sedangkan jika LED hijau menyala maka Sensor Infrared Proximity E18-D80NK akan aktif dan objek pejalan kaki tidak dapat menyeberang.
3	ESP32-Cam	ESP32-Cam dapat menangkap gambar objek pejalan kaki pada saat terdeteksi pelanggaran penyeberangan dan gambar dikirim melalui Bot Telegram.	Berhasil	ESP32-Cam dapat menangkap gambar pada saat objek pejalan kaki terdeteksi oleh Sensor Infrared Proximity E18-D80NK dalam keadaan ON. Hasil tangkapan gambar akan ditampilkan melalui Bot Telegram.
4	*Water Spray*	*Water spray* dapat menyemprotkan air pada saat terdeteksinya objek pejalan kaki melanggar penyeberangan pada *zebra cross*.	Berhasil	Pada saat lampu lalu lintas dalam keadaan berwarna merah dan kuning, maka *water spray* dalam keadaan OFF. Lalu pada saat lampu lalu lintas dalam keadaan berwarna hijau, maka *water spray* dalam keadaan ON.

Pengujian interaksi ESP32-Cam dengan Bot Telegram, tujuannya untuk mengetahui kecepatan ESP32-Cam dalam menangkap gambar suatu pelanggaran kemudian dikirimkan melalui Bot Telegram pada saat terjadinya pelanggaran penyeberangan pada *zebra cross*. Tabel 3 merupakan pengujian interaksi ESP32-Cam dengan Bot Telegram. Berdasarkan Tabel 3 diketahui ESP32-Cam berhasil menangkap gambar dan Bot Telegram berhasil mengirim gambar pelanggaran pada *zebra cross*. Pengujian *delay* pengiriman gambar ESP32-Cam dengan Bot Telegram menggunakan *stopwatch* dengan percobaan sebanyak 10 kali pelanggaran. Dari hasil perhitungan tersebut, didapatkan nilai terendah pada *delay* pengiriman gambar Bot Telegram 4,44 detik dan didapatkan niai tertinggi pada *delay* pengiriman gambar Bot Telegram 4,86 detik. Kemudian diketahui untuk rata-rata *delay* keseluruhan pengiriman gambar Bot Telegram adalah

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

4,65 detik. Kecepatan pengiriman gambar cepat atau lambat bergantung pada kondisi jaringan internet.

Tabel 2 Pengujian *Delay* Respon *Water Spray* dan *Buzzer*

No.	Banyaknya Pelanggaran	*Delay* Respon *Water Spray* (detik)	*Delay* Respon *Buzzer* (detik)
1	Pelanggaran 1	0,80	0,79
2	Pelanggaran 2	0,82	0,82
3	Pelanggaran 3	0,84	0,84
4	Pelanggaran 4	0,90	0,88
5	Pelanggaran 5	0,80	0,73
6	Pelanggaran 6	0,83	0,83
7	Pelanggaran 7	0,87	0,85
8	Pelanggaran 8	0,90	0,90
9	Pelanggaran 9	0,99	0,90
10	Pelanggaran 10	1,01	0,93
Rata-rata *Delay*		**0,87**	**0,84**
		0,85	

Tabel 3 Pengujian Interaksi ESP32-Cam Dengan Bot Telegram

Banyaknya Pelanggaran	ESP2-Cam	Bot Telegram	*Delay* Mengirim Gambar (detik)
Pelanggaran 1	Menangkap Gambar	Gambar Terkirim	4,71
Pelanggaran 2	Menangkap Gambar	Gambar Terkirim	4,61
Pelanggaran 3	Menangkap Gambar	Gambar Terkirim	4,44
Pelanggaran 4	Menangkap Gambar	Gambar Terkirim	4,84
Pelanggaran 5	Menangkap Gambar	Gambar Terkirim	4,48
Pelanggaran 6	Menangkap Gambar	Gambar Terkirim	4,56
Pelanggaran 7	Menangkap Gambar	Gambar Terkirim	4,64
Pelanggaran 8	Menangkap Gambar	Gambar Terkirim	4,78
Pelanggaran 9	Menangkap Gambar	Gambar Terkirim	4,66
Pelanggaran 10	Menangkap Gambar	Gambar Terkirim	4,86
Rata-rata *Delay*			**4,65**

Pengujian untuk akurasi dilakukan menggunakan *confusion matrix* dengan acuan 4 (empat) keadaan umum yang digunakan sebagai tolak ukur dalam pengujian ini, yaitu *True Positive* (TP) adalah keadaan di saat sistem mendeteksi adanya pelanggaran penyeberangan ketika terjadi pelanggaran penyeberangan, *True Negative* (TN) adalah keadaan di saat sistem tidak mendeteksi pelanggaran penyeberangan ketika tidak terjadi pelanggaran penyeberangan, *False Positive* (FP) adalah keadaan di saat sistem mendeteksi pelanggaran penyeberangan ketika tidak terjadi pelanggaran penyeberangan, dan *False Negative* (FN) adalah keadaan di saat sistem tidak mendeteksi pelanggaran penyeberangan ketika terjadi pelanggaran penyeberangan.

Dilakukan sebanyak 30 kali percobaan, sistem memprediksi adanya 17 pejalan kaki yang terdeteksi melanggar penyeberangan dan yang diprediksi tidak terdeteksi sebanyak 2 pejalan

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

kaki. Dan dari 13 pejalan kaki yang tidak terdeteksi pelanggaran penyeberangan, model memprediksi ada 3 pejalan kaki yang diprediksi melanggar penyeberangan. Pernyataan tersebut dapat terlihat pada Tabel 4.

Tabel 4 *Confusion Matrix*

N = 30	Aktual Positif (+)	Aktual Negatif (-)
Prediksi Positif (+)	15 TP	3 FP
Prediksi Negatif (-)	2 FN	10 TN

Berdasarkan Tabel 4 telah dilakukan sebanyak 30 percobaan maka didapatkan aktual positif dan prediksi positif sebanyak 15 TP, aktual negatif dan prediksi positif 3 TP, aktual positif dan prediksi negatif 2 FN, dan aktual negatif dan prediksi negatif 10 TN. Setelah diketahui jumlah dari data yang benar diprediksi maupun yang salah diprediksi dapat dilakukan kalkulasi untuk mengetahui *accuracy*, *precision*, dan *recall* dari seluruh data tersebut. Setelah diketahui jumlah data dari aktual positif, aktual negatif, prediksi positif, dan prediksi negatif dapat dilakukan perhitungan pada Tabel 5 untuk mengetahui *accuracy*, *precision*, dan *recall* dari pendeteksian pelanggaran penyeberangan pada *zebra cross*.

Tabel 5 Hasil Perhitungan *Confusion Matrix*

Aktual dan Prediksi (%)	
Accuracy	0,8333
Precision	0,8333
Recall	0,8823

Pengamatan yang dilakukan dari Tabel 5 hasil perhitungan *confusion matrix* untuk pendeteksian pelanggaran penyeberangan pada *zebra cross* diperoleh nilai *accuracy* sebesar 83,33%, kemudian diperoleh nilai *precision* sebesar 83,33%, dan diperoleh nilai *recall* sebesar 88,23%. Berdasarkan hasil yang telah diperoleh dapat disimpulkan bahwa nilai *accuracy*, *precision*, dan *recall* sudah baik, sehingga sistem deteksi pelanggaran penyeberangan pada *zebra cross* dapat melakukan pendeteksian dengan baik.

4. KESIMPULAN

Berdasarkan pada hasil pembahasan dan hasil pengujian, sistem dapat mendeteksi pelanggaran penyeberangan dan bekerja dengan baik apabila terhubung dengan jaringan internet yang stabil. Sensor Infrared Proximity E18-D80NK dapat mendeteksi objek pada saat lampu lalu lintas berwarna hijau, *water spray* dapat menyemprotkan air, dan *buzzer* dapat mengeluarkan suara pada saat terjadinya pelanggaran penyeberangan. Selanjutnya kamera dapat menangkap gambar dan dikirim ke Bot Telegram. Dari hasil perhitungan menggunakan pengujian *confusion matrix* didapatkan nilai *accuracy* sebesar 83,33%, nilai *precision* sebesar 83,33%, dan nilai *recall* sebesar 88,23%. Pada sistem ini masih terdapat kekurangan sehingga diperlukan pengembangan dan penelitian lebih lanjut untuk penyempurnaannya, seperti dapat memfokuskan pada teknologi baru untuk meningkatkan akurasi pelanggaran penyeberangan seperti menggunakan kamera CCTV yang memiliki fitur AI (*Artificial Intelligence*) guna mengidentifikasi pada pendeteksian objek pejalan kaki saja. Penelitian ini dapat dikembangkan menggunakan dua atau lebih sampel sensor untuk mendeteksi objek dari berlawanan arah dengan kondisi cuaca yang berbeda-beda seperti cerah dan hujan sehingga menambah keakuratan dari sistem ini. Bot Whatsapp dapat dikembangkan lebih lanjut karena hampir semua orang menggunakan aplikasi ini.

DAFTAR PUSTAKA

Al Khaledi, M. T., Nasri, N., & Hanafi, H. (2022). Rancang Bangun Sistem Rumah Pintar Menggunakan Platform Google Firebase Berbasis IoT (Internet of Things). *Jurnal TEKTRO*, *6*(2), 194–202. https://doi.org/10.30811/TEKTRO.V6I2.3732

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

Alamsyah, I. E. (2019). *Ingat, Berhenti di Zebra Cross adalah Pelanggaran!* Republika Online. https://www.republika.co.id/berita/pvk34l349/ingat-berhenti-di-emzebra-crossem-adalah-pelanggaran

Damayanti, F. P., Riyadi, M. A., & Andromeda, T. (2019). Perancangan Traffic Light Menggunakan Modul Field Programmable Gate Array Xilinx Nexys 3. *Transient: Jurnal Ilmiah Teknik Elektro*, *7*(2), 678–685. https://doi.org/10.14710/TRANSIENT.V7I2.678-685

Düntsch, I., & Gediga, G. (2019). Confusion Matrices and Rough Set Data Analysis. *Journal of Physics: Conference Series*, *1229*(1), 012055. https://doi.org/10.1088/1742-6596/1229/1/012055

Embrianto, S. E., & Sulistyowati, E. (2020). Pengawasan Terhadap Pejalan Kaki yang Tidak Menyeberang di Tempat Penyeberangan Pejalan Kaki di Kota Surabaya. *NOVUM: JURNAL HUKUM*, *7*(3). https://doi.org/10.2674/NOVUM.V7I3.32321

Hasanah, U., Subito, M., & Indrajaya, M. A. (2021). Rancang Bangun Prototype Sistem Pendeteksi Pelanggaran pada Zebra Cross di Lampu Lalu Lintas Berbasis Arduino. *Foristek*, *11*(1), 1–7. https://doi.org/10.54757/fs.v11i1.31

IOTkece. (2021). *Apa itu ESP32? Spesifikasi ESP32*. IOTkece. https://iotkece.com/apa-itu-esp32-spesifikasi-esp32/

Mahmuddin, A., Febryan, A., Adriani, A., & Rahmania, R. (2023). Rancang Bangun Sistem Keamanan Rumah Berbasis Telegram Menggunakan ESP 32 Cam. *VERTEX ELEKTRO*, *15*(1), 64–71. https://journal.unismuh.ac.id/index.php/vertex/article/view/10246

Nugroho, Y. A. (2018). Keamanan Dan Kenyamanan Trotoar Di Taman Tingkir, Kota Salatiga. *Mintakat: Jurnal Arsitektur*, *19*(1), 35–48. https://doi.org/10.26905/mintakat.v19i1.1440

Prabowo, R. R., Kusnadi, K., & Subagio, R. T. (2020). Sistem Monitoring dan Pemberian Pakan Otomatis pada Budidaya Ikan Menggunakan Wemos dengan Konsep Internet of Things (IoT). *Jurnal Digit*, *10*(2), 185. https://doi.org/10.51920/jd.v10i2.169

Pressman, R. S., & Maxim, B. R. (2019). *Software Engineering: A Practitioner's Approach* (9th ed.). McGraw Hill. https://www.researchgate.net/publication/336251012_Software_Engineering_A_Practitioner's_Approach_9th_Edition

Rahmawati, Y., Simanjuntak, I. U. V., & Simorangkir, R. B. (2022). Rancang Bangun Purwarupa Sistem Peringatan Pengendara Pelanggar Zebra Cross Berbasis Mikrokontroller ESP-32 CAM. *Jambura Journal of Electrical and Electronics Engineering*, *4*(2), 189–195. https://doi.org/10.37905/jjeee.v4i2.14499

Saputra, A. F., & Darujati, C. (2020). Sistem Presensi Mahasiswa Berbasis Realtime Kamera Metode Klasifikasi Haar. *Jurnal Teknik Elektro Dan Komputer*, *9*(3), 137–144. https://doi.org/10.35793/JTEK.V9I3.29488

Ulum, Moch. B., Lutfi, Moch., & Faizin, A. (2022). Otomatisasi Pompa Air Menggunakan NodeMCU ESP8266 Berbasis Internet of Things (IoT). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *6*(1), 86–93. https://doi.org/10.36040/jati.v6i1.4583

Wahyudi, M. I., & Aziz, R. A. (2022). Keran Air Wudhu Otomatis Menggunakan Sensor Infrared Sebagai Upaya Meminimalisasi Pemborosan Air. *Journal of Applied Computer Science and Technology*, *3*(1), 151–156. https://doi.org/10.52158/jacost.v3i1.296

Wicaksono, A., Purnomo, M. H., & Yuniarno, E. M. (2021). Deteksi Pejalan Kaki pada Zebra Cross untuk Peringatan Dini Pengendara Mobil Menggunakan Mask R-CNN. *Jurnal Teknik ITS*, *10*(2), A497–A503. https://doi.org/10.12962/j23373539.v10i2.80219

Widyaningsih, N., & Daniel, O. (2019). Analisis Karakteristik dan Perilaku Penyeberangan Orang pada Fasilitas Penyeberangan Zebra Cross dan Pelican Cross (Studi Kasus Ruas Jalan M. H. Thamrin). *Jurnal Pengembangan Rekayasa Dan Teknologi*, *15*(1), 27–32. https://doi.org/10.26623/jprt.v15i1.1486

Yolanda, M., Rahmat, B., & Hertina, S. N. (2021). Pendeteksi Pelanggaran Penyeberang Jalan pada Zebra Cross Berbasis Internet of Things. *EProceedings of Engineering*, *8*(5), 5211–5220. https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15850

Artikel ini didistribusikan mengikuti lisensi Atribusi-NonKomersial CC BY-NC sebagaimana tercantum pada https://creativecommons.org/licenses/by-nc/4.0/.

9 772527 583007