KOMPARASI ALGORITMA SUPPORT VECTOR MACHINE DAN NAIVE BAYES PADA ANALISIS SENTIMEN DATA TWITTER

Anas Shoffy Muammar[1*)], Annisa Al Mawy[2]
Teknik Komputer
*) annisaalmawy09@gmail.com

Abstrak

Twitter merupakan media sosial yang banyak digunakan. Media sosial twitter dapat digunakan untuk mengekspresikan pikiran dan pendapat sesorang tentang objek tersebut. Ini memberikan peluang besar untuk menggunakan sumber data untuk analisis sentimen. Ada banyak algoritma untuk melakukan analisis sentimen, seperti Support Vector Machine (SVM) dan Naive Bayes (NB). Peneliti tertarik menggunakan metode Support Vector Machine dan Naïve Bayes untuk mengklasifikasikan data, karena ada banyak pendapat tentang kinerja kedua metode tersebut. Data yang digunakan dalam penelitian ini adalah data opini masyarakat mengenai kebijakan vaksinasi Covid-19. Proses klasifikasi pertama dilakukan dengan metode Support Vector Machine dengan kernel yang berbeda. Setelah didapatkan hasil akurasi terbaik, hasil akurasi tersebut dibandingkan dengan nilai akurasi dari hasil naive bayes classifier.

Kata Kunci: Support Vector Machine, Naïve Bayes. Analisis Sentimen, Covid-19

PENDAHULUAN

Banyak orang kini menyuarakan pendapatnya melalui media sosial (Rahman Isnain et al., 2021);(Susanto & Puspaningrum, 2019);(Firmansyah et al., 2017). Opini yang dikomunikasikan melalui media sosial lebih interaktif daripada media cetak. Salah satu media sosial yang paling banyak digunakan saat ini adalah Twitter (Alita, Sari, et al., 2021);(Alita, 2021);(Styawati, Hendrastuty, et al., 2021). Menurut We are Social sources pada tahun 2020, di media sosial, Twitter adalah kategori media sosial kelima yang paling banyak digunakan setelah Youtube, Whatsapp, Facebook dan Instagram, dengan 56% pengguna. Hal ini menunjukkan bahwa ada peluang besar untuk menggunakan sumber data untuk menganalisis perasaan orang terhadap objek. Ada banyak algoritma untuk melakukan analisis sentimen, seperti Support Vector Machines (SVM) dan Naive Bayes (NB). Klasifikasi data Twitter menggunakan metode SVM dan NB menghasilkan akurasi tertinggi di antara kedua metode tersebut, yaitu metode NB, dengan nilai akurasi sebesar 94%. Berdasarkan penelitian sebelumnya, penelitian ini akan membandingkan nilai akurasi hasil klasifikasi data Twitter. Data Twitter yang digunakan adalah data opini publik terkait vaksinasi Covid-19 (Rachman & Pramana, 2020);(Arpiansah et al., 2021). Data ini dipilih karena vaksinasi Covid-19 sempat menjadi trending topic di Twitter.

Tujuan dari penelitian ini adalah untuk membandingkan akurasi metode SVM dengan berbagai kernel (Hendrastuty et al., 2021);(Isnain et al., 2021). Selain itu, penelitian ini juga bertujuan untuk membandingkan akurasi metode SVM dengan metode NB. Perbandingan ini dilakukan untuk mengetahui performansi dari kedua metode klasifikasi tersebut.

KAJIAN PUSTAKA

Data Prapemrosesan

Tujuan dari data preprocessing adalah untuk membersihkan data, integrasi data, transformasi data, dan reduksi data. Preprocessing dalam penelitian ini menggunakan lima teknik yaitu cleansing, tokenization, case folding, stopword removal, dan stemming (Indrayuni, 2019);(Wahyono et al., 2021).

1. Pembersihan

Pembersihan data merupakan kegiatan analisis kualitas data. Hal ini dapat dilakukan dengan mengoreksi, mengubah atau menghapus data dalam database atau dalam format file yang salah yang dianggap tidak perlu, tidak lengkap atau tidak akurat (Giovani et al., 2020). Contoh: "Program imunisasi pemerintah kacau dengan orang-orang yang mengantri untuk membatalkan vaksin yang gagal" hingga "Program imunisasi pemerintah sedang mengantre untuk membuat vaksin yang gagal Itu kacau di mana pun saya berada."

2. Tokenisasi

Tokenisasi digunakan untuk memecah komentar menjadi kata-kata. Proses tokenization dilakukan dengan melihat setiap ruang di dalam komentar, setelah itu komentar dapat dibatasi berdasarkan ruang tersebut (D. Ariyanti & Iswardani, 2020);(Hendrastuty, 2021). Misalnya, "Program imunisasi gagal di mana saja dalam antrian untuk membuat vaksin gratis" menjadi "[program, vaksin, pemerintah, kacau, di mana, di mana, antri, dibuat, Dibeli, ditulis, vaksin gagal akan berada.

3. Kasus lipat

Case folding adalah proses memodifikasi catatan teks untuk konsistensi . Kapitalisasi dilakukan dengan mengubah teks ke bentuk kanoniknya (biasanya huruf kecil, juga dikenal sebagai huruf kecil (Isnain et al., n.d.) . Misalnya, "Program vaksinasi pemerintah kacau di mana antrian untuk vaksin gratis gagal" hingga "Program vaksin gagal di sepanjang antrian vaksin gagal".

4. Stopword

Stopwords adalah proses menghilangkan kata-kata yang terdapat dalam daftar stopword(Purwarianti, 2014). Stopwords adalah kata umum yang banyak muncul dan memiliki fungsi tetapi tidak memiliki arti. Contoh stopwords termasuk "apa" dan "atau".

5. Stemming

Data stemming adalah proses penyaringan kata yang mengandung kata penghubung, kata ganti, dan kata depan menjadi kata yang lebih sederhana dengan menghilangkan awalan atau akhiran (Abidin et al., 2021);(Borman et al., 2018). Misalnya, "Jika program vaksin gagal di mana pun dalam antrian, vaksin gagal" menjadi "Jika program vaksin gagal di mana pun dalam antrian, vaksin gagal".

Validasi Data

Validasi data adalah langkah verifikasi untuk memastikan bahwa data memenuhi standar yang ditetapkan, memastikan bahwa data yang masuk ke database diketahui dan dapat dijelaskan sumber dan keakuratan datanya (Nabila et al., 2021). Teknik validasi data yang digunakan adalah validasi silang K-Fold (Nasution & Hayaty, 2019);(Sulistiani et al., 2019). K-Fols cross validation merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan suatu system. Penelitian ini menggunakan 10 lipatan.

Data Latih

Data latih adalah bagian dari kumpulan data yang Anda latih untuk membuat prediksi sesuai dengan tujuan tertentu atau menjalankan fungsi algoritma lainnya (Kurniawan & Susanto, 2019);(Neneng, Puspaningrum, Lestari, et al., 2021). Kami memberikan petunjuk melalui algoritme sehingga mesin yang kami latih dapat menemukan korelasinya sendiri (Styawati, Andi Nurkholis, et al., 2021);(Setiawan & Pasha, 2020).

Data Uji

Data uji digunakan untuk menentukan kinerja algoritme yang dilatih sebelumnya saat baru, belum pernah terlihat sebelum data ditemukan (L. Ariyanti et al., 2020);(Sulistiyawati & Supriyanto, 2021). Ini biasanya disebut generalisasi. Hasil penelitian ini dapat merujuk pada hasil pelatihan sebagai model.

Klasifikasi Support Vector Machine (SVM)

Support vector machine (SVM) adalah salah satu algoritma pembelajaran mesin terawasi yang berkinerja baik dalam mengklasifikasikan data. SVM juga dikenal sebagai pengklasifikasi linier berdasarkan prinsip maksimalisasi margin (Neneng et al., 2016);(Styawati et al., 2022). SVM memanfaatkan hyperplanes secara optimal untuk mengklasifikasikan data menjadi dua kelompok dalam ruang dimensi tinggi. Jarak adalah jarak antara hyperplane dengan data terdekat untuk setiap kelas (Fikri et al., 2020). Data terdekat ini disebut support vector. Hyperplane adalah pemisah optimal antara dua kelas yang telah ditentukan. Prinsip dasar SVM adalah pengklasifikasi linier, yang kemudian dikembangkan untuk mengatasi masalah nonlinier dengan memasukkan konsep trik kernel ke dalam ruang kerja berdimensi tinggi. Kernel SVM yang digunakan dalam penelitian ini adalah kernel linear, radial basis function (RBF), dan polynomial (Aldino et al., 2021);(Alita, Putra, et al., 2021). Metode SVM memiliki konsep utama dalam mengklasifikasikan data. Hyperplane optimal diperoleh dengan memaksimalkan garis lintang dari support vektor.

Klasifikasi Naive Bayes

Naive Bayes (NB) adalah metode klasifikasi yang dapat memprediksi kemungkinan kelas dan mengambil keputusan berdasarkan data pelatihan (Sengkey et al., 2020). Naïve bayes dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Handoko & Neneng, 2021);(Gandhi et al., 2021);(Redy Susanto et al., 2021). Naïve Bayes memiliki keunggulan seperti kesederhanaan, kecepatan, dan akurasi yang tinggi ketika diterapkan pada data dalam jumlah besar. Secara umum rumus klasifikasi Naïve Bayes dapat dilihat pada rumus (Ferdiana, 2020).

METODE

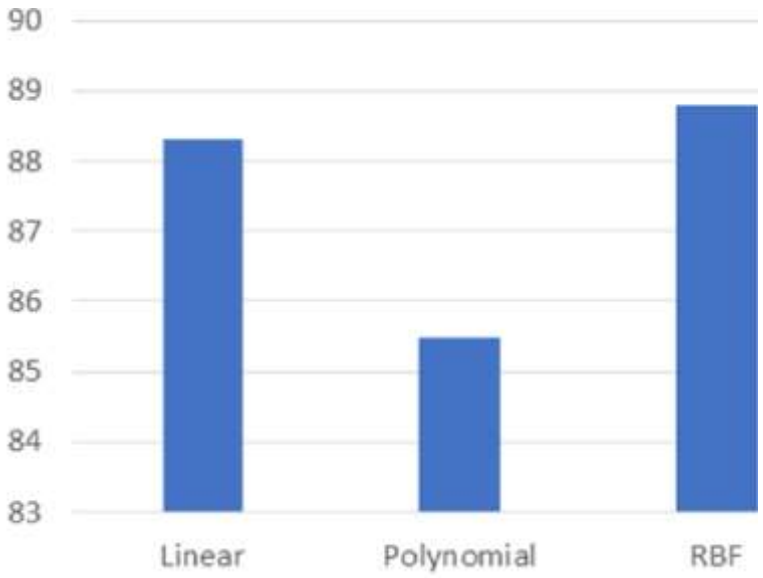
Pemodelan SVM

Kernel linier dalam metode SVM berfungsi untuk memisahkan data secara linier . Source code untuk proses klasifikasi menggunakan kernel Linear. Baris pertama akan membuat variabel clf yang berisi SVC(Support Vector Classifier) dengan kernel linier dan C=2.33. Hasil klasifikasi menggunakan metode linear kernel SVM dengan mencoba berbagai nilai C.

Kernel Polynomial adalah fungsi kernel yang digunakan ketika data tidak dapat dipisahkan secara linier (Neneng, Puspaningrum, & Aldino, 2021). Source code untuk proses klasifikasi

menggunakan kernel Polynomial. Hasil klasifikasi menggunakan metode Polynomial kernel SVM dengan mencoba berbagai nilai C, akurasi tertinggi adalah 85,5. Akurasi diperoleh dari penggunaan parameter C=2.33, C=2.25, C=2.13.

Kernel RBF berfungsi untuk memisahkan data dengan dimensi yang lebih tinggi. Source code untuk proses klasifikasi menggunakan kernel RBF. Hasil klasifikasi menggunakan metode SVM kernel RBF dengan mencoba berbagai nilai C, didapat akurasi tertinggi sebesar 88,8. Akurasi diperoleh dari penggunaan parameter C = 2.13 dan gamma = 0.50.



Gambar 1 Perbandingan Kernel SVM

Berdasarkan gambar 1 diatas dapat dilihat bahwa akurasi paling tinggi diperoleh dari metode SVM dengan kernel RBF. Kernel RBF mendapatkan akurasi yang lebih tinggi dibandingkan dengan kernel lainnya.

Pemodelan Naive Bayes

Metode Naive Bayes merupakan metode klasifikasi teks berdasarkan probabilitas kata kunci dalam membandingkan dokumen pelatihan dan dokumen pengujian (Darwis et al., 2021). Keduanya dibandingkan melalui beberapa tahap persamaan, yang pada akhirnya menghasilkan probabilitas tertinggi untuk ditetapkan sebagai kategori dokumen baru. Berikut source code untuk proses klasifikasi menggunakan Naive Bayes

```
from sklearn.naive bayes import GaussianNB
```

```
modeInb = GaussianNB()
nbtrain = modelnb.fit(x_train, y_train)
y_pred = nbtrain.predict(x_test)
nbtrain.predict_proba(x_test)
from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred)
```

Akurasi yang diperoleh menggunakan metode naive bayes adalah sebesar 82,51%.

KESIMPULAN

Hasil penelitian ini menunjukkan bahwa SVM dengan kernel RBF memberikan akurasi terbaik dibandingkan dengan kernel linier dan polinomial. Hal ini karena saat melakukan assign data, kernel RBF memperhitungkan nilai yang digunakan untuk mencari nilai optimal di setiap kumpulan data (gamma). Hasil yang diperoleh membandingkan akurasi RBF kernel SVM dan NB, nilai akurasi tertinggi berasal dari metode RBF kernel SVM, yaitu 88,8%. Ini karena kami menggunakan kumpulan data yang tidak terlalu besar. Selain itu, NB menggunakan nilai probabilitas dalam proses klasifikasi data.

REFERENSI

Abidin, Z., Wijaya, A., & Pasha, D. (2021). Aplikasi Stemming Kata Bahasa Lampung Dialek Api Menggunakan Pendekatan Brute-Force dan Pemograman C. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(1), 1–8.

Aldino, A. A., Saputra, A., & Nurkholis, A. (2021). *Application of Support Vector Machine (SVM) Algorithm in Classification of Low-Cape Communities in Lampung Timur*. *3*(3), 325–330. https://doi.org/10.47065/bits.v3i3.1041

Alita, D. (2021). Multiclass SVM Algorithm for Sarcasm Text in Twitter. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *8*(1), 118–128. https://doi.org/10.35957/jatisi.v8i1.646

Alita, D., Putra, A. D., & Darwis, D. (2021). Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(3), 1–5.

Alita, D., Sari, I., Isnain, A. R., & Styawati, S. (2021). Penerapan Naïve Bayes Classifier Untuk Pendukung Keputusan Penerima Beasiswa. *Jurnal Data Mining Dan Sistem Informasi*, *2*(1), 17–23.

Ariyanti, D., & Iswardani, K. (2020). Teks Mining untuk Klasifikasi Keluhan Masyarakat Pada Pemkot Probolinggo Menggunakan Algoritma Naïve Bayes. *Jurnal IKRA-ITH Informatika*, *4*(3), 125–132.

Ariyanti, L., Najib, M., Satria, D., & Alita, D. (2020). Sistem Informasi Akademik Dan

Administrasi Dengan Metode Extreme Programming Pada Lembaga Kursus Dan Pelatihan. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *1*(1), 90–96. http://jim.teknokrat.ac.id/index.php/sisteminformasi

Arpiansah, R., Fernando, Y., & Fakhrurozi, J. (2021). Game Edukasi VR Pengenalan Dan Pencegahan Virus Covid-19 Menggunakan Metode MDLC Untuk Anak Usia Dini. *Jurnal Teknologi Dan Sistem Informasi*, *2*(2), 88–93.

Borman, R. I., Putra, Y. P., Fernando, Y., Kurniawan, D. E., Prasetyawan, P., & Ahmad, I. (2018). Designing an Android-based Space Travel Application Trough Virtual Reality for Teaching Media. *2018 International Conference on Applied Engineering (ICAE)*, 1–5.

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, *15*(1), 131–145.

Ferdiana, R. (2020). A Systematic Literature Review of Intrusion Detection System for Network Security: Research Trends, Datasets and Methods. *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, 1–6.

Fikri, M. I., Sabrila, T. S., & Azhar, Y. (2020). Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter. *Smatika Jurnal*, *10*(02), 71–76. https://doi.org/10.32664/smatika.v10i02.455

Firmansyah, M. A., Karlinah, S., & Sumartias, S. (2017). Kampanye Pilpres 2014 dalam Konstruksi Akun Twitter Pendukung Capres. *Jurnal The Messenger*, *9*(1), 79. https://doi.org/10.26623/themessenger.v9i1.430

Gandhi, B. S., Megawaty, D. A., & Alita, D. (2021). Aplikasi Monitoring Dan Penentuan Peringkat Kelas Menggunakan Naïve Bayes Classifier. *Jurnal Informatika Dan Rekayasa Perangkat Lunak*, *2*(1), 54–63.

Giovani, A. P., Ardiansyah, A., Haryanti, T., Kurniawati, L., & Gata, W. (2020). Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi. *Jurnal Teknoinfo*, *14*(2), 115. https://doi.org/10.33365/jti.v14i2.679

Handoko, M. R., & Neneng, N. (2021). SISTEM PAKAR DIAGNOSA PENYAKIT SELAMA KEHAMILAN MENGGUNAKAN METODE NAIVE BAYES BERBASIS WEB. *Jurnal Teknologi Dan Sistem Informasi*, *2*(1), 50–58.

Hendrastuty, N. (2021). *Text Summarization in Multi Document Using Genetic Algorithm*. *15*(4), 327–338.

Hendrastuty, N., Rahman Isnain, A., & Yanti Rahmadhani, A. (2021). *Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine*. *6*(3), 150–155. http://situs.com

Indrayuni, E. (2019). Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes. *Jurnal Khatulistiwa Informatika*, *7*(1), 29–36. https://doi.org/10.31294/jki.v7i1.1

Isnain, A. R., Hendrastuty, N., Andraini, L., Studi, P., Informasi, S., Indonesia, U. T., Informatika, P. S., Indonesia, U. T., Studi, P., Komputer, T., Indonesia, U. T., & Lampung, K. B. (2021). *Comparison of Support Vector Machine and Naïve Bayes on Twitter Data Sentiment Analysis*. *6*(1), 56–60.

Isnain, A. R., Marga, N. S., & Alita, D. (n.d.). Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(1), 55–64.

Kurniawan, I., & Susanto, A. (2019). Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019. *Eksplora Informatika*, *9*(1), 1–10. https://doi.org/10.30864/eksplora.v9i1.237

Nabila, Z., Rahman Isnain, A., & Abidin, Z. (2021). Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(2), 100. http://jim.teknokrat.ac.id/index.php/JTSI

Nasution, M. R. A., & Hayaty, M. (2019). Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter. *Jurnal Informatika*, *6*(2), 226–235. https://doi.org/10.31311/ji.v6i2.5129

Neneng, N., Adi, K., & Isnanto, R. (2016). Support Vector Machine Untuk Klasifikasi Citra Jenis Daging Berdasarkan Tekstur Menggunakan Ekstraksi Ciri Gray Level Co-Occurrence Matrices (GLCM). *JSINBIS (Jurnal Sistem Informasi Bisnis)*, *6*(1), 1–10.

Neneng, N., Puspaningrum, A. S., & Aldino, A. A. (2021). Perbandingan Hasil Klasifikasi Jenis Daging Menggunakan Ekstraksi Ciri Tekstur Gray Level Co-occurrence Matrices (GLCM) Dan Local Binary Pattern (LBP). *SMATIKA JURNAL*, *11*(01), 48–52.

Neneng, N., Puspaningrum, A. S., Lestari, F., & Pratiwi, D. (2021). SMA Tunas Mekar Indonesia Tangguh Bencana. *Jurnal Pengabdian Masyarakat Indonesia*, *1*(6), 335–342. https://doi.org/10.52436/1.jpmi.61

Purwarianti, A. (2014). Rule based approach for text segmentation on Indonesian news article using named entity distribution. *2014 International Conference on Data and Software Engineering (ICODSE)*, 1–5.

Rachman, F. F., & Pramana, S. (2020). *Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter*. *8*(2), 100–109.

Rahman Isnain, A., Indra Sakti, A., Alita, D., & Satya Marga, N. (2021). Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma Svm. *Jdmsi*, *2*(1), 31–37. https://t.co/NfhnfMjtXw

Redy Susanto, E., Admi Syarif, A. S., Muludi, K., & Wantoro, A. (2021). *Peer Review: Implementation of Fuzzy-based Model for Prediction of Thalassemia Diseases*.

Sengkey, D. F., Kambey, F. D., Lengkong, S. P., Joshua, S. R., & Kainde, H. V. F. (2020). Pemanfaatan Platform Pemrograman Daring dalam Pembelajaran Probabilitas dan Statistika di Masa Pandemi CoVID-19. *Jurnal Informatika*, *15*(4), 217–224.

Setiawan, A., & Pasha, D. (2020). Sistem Pengolahan Data Penilaian Berbasis Web Menggunakan Metode Pieces (Studi Kasus : Badan Pengembangan Sumber Daya Manusia Provinsi Lampung). *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *1*(1), 97–104. http://jim.teknokrat.ac.id/index.php/sisteminformasi

Styawati, Andi Nurkholis, Zaenal Abidin, & Heni Sulistiani. (2021). Optimasi Parameter Support Vector Machine Berbasis Algoritma Firefly Pada Data Opini Film. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(5), 904–910. https://doi.org/10.29207/resti.v5i5.3380

Styawati, S., Hendrastuty, N., & Isnain, A. R. (2021). Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine. *Jurnal Informatika: Jurnal Pengembangan IT*, *6*(3), 150–155.

Styawati, S., Nurkholis, A., Aldino, A. A., Samsugi, S., Suryati, E., & Cahyono, R. P. (2022). Sentiment Analysis on Online Transportation Reviews Using Word2Vec Text Embedding Model Feature Extraction and Support Vector Machine (SVM) Algorithm. *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, 163–167.

Sulistiani, H., Muludi, K., & Syarif, A. (2019). Implementation of Dynamic Mutual Information and Support Vector Machine for Customer Loyalty Classification. *Journal of Physics: Conference Series*, *1338*(1). https://doi.org/10.1088/1742-6596/1338/1/012050

Sulistiyawati, A., & Supriyanto, E. (2021). Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan. *Jurnal Tekno Kompak*, *15*(2), 25. https://doi.org/10.33365/jtk.v15i2.1162

Susanto, E. R., & Puspaningrum, A. S. (2019). *Rancang Bangun Rekomendasi Penerima Bantuan Sosial Berdasarkan Data Kesejahteraan Rakyat*. *15*(1), 1–12.

Wahyono, Wibowo, M. E., Ashari, A., & Putra, M. P. K. (2021). Improvement of Deep Learning-based Human Detection using Dynamic Thresholding for Intelligent Surveillance System. *International Journal of Advanced Computer Science and Applications*, *12*(10), 472–477. https://doi.org/10.14569/IJACSA.2021.0121053