Journal of Science and Social Research ISSN 2615 – 4307 (Print)
April 2026, IX (2): 2062 – 2067 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

PERBANDINGAN KINERJA ALGORITMA LOGISTIC REGRESSION DAN SUPPORT VECTOR MACHINE DALAM PREDIKSI RISIKO PENGGUNAAN NARKOBA

Revi Afriani Siringo Ringo[1], Maykel Oliver Sitohang[2], Dora Etimanta br. Ginting[3], Zimmy Silalahi[4]
Universitas Prima Indonesia, Medan
e-mail: [1]reviafriani17@gmail.com, [2]maykeloliverrsitohang@gmail.com,
[3]doraetimanta3@gmail.com, [4]zimmys766@gmail.com

Abstract: *Drug abuse constitutes a global public health issue that significantly affects individuals as well as social systems. Early detection of high-risk individuals represents a strategic approach in prevention and intervention efforts. This study aims to compare the performance of the Logistic Regression and Support Vector Machine algorithms in predicting the risk of drug use using the Drug Consumption (Quantified) Dataset from the UCI Machine Learning Repository, consisting of 1,885 observations. The research process includes data preprocessing stages such as data cleaning and feature normalization, feature selection using Recursive Feature Elimination, and handling class imbalance through the SMOTE method. Model evaluation is conducted using various classification metrics, including accuracy, precision, recall, F1-score, AUC-ROC, Matthews Correlation Coefficient, and G-Mean, as well as statistical testing to assess performance differences between the two algorithms. This study is expected to contribute to the development of a machine learning–based early screening system to support decision-making in the prevention of drug abuse.*

Keywords*: Drug Abuse, Logistic Regression, Support Vector Machine, Classification, Machine Learning.*

Abstrak: Penyalahgunaan narkoba merupakan permasalahan kesehatan masyarakat global yang berdampak signifikan terhadap individu maupun sistem sosial. Deteksi dini terhadap individu berisiko tinggi menjadi langkah strategis dalam upaya pencegahan dan intervensi. Penelitian ini bertujuan untuk membandingkan kinerja algoritma Logistic Regression dan Support Vector Machine dalam memprediksi risiko penggunaan narkoba menggunakan Drug Consumption (Quantified) Dataset dari UCI Machine Learning Repository yang berjumlah 1.885 observasi. Proses penelitian meliputi tahap pra-pemrosesan data berupa pembersihan dan normalisasi fitur, seleksi fitur menggunakan Recursive Feature Elimination, serta penanganan ketidakseimbangan kelas dengan metode SMOTE. Evaluasi model dirancang menggunakan berbagai metrik klasifikasi seperti accuracy, precision, recall, F1-score, AUC-ROC, Matthews Correlation Coefficient, dan G-Mean, serta pengujian statistik untuk menilai perbedaan performa kedua algoritma. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem skrining dini berbasis machine learning sebagai pendukung pengambilan keputusan dalam pencegahan penyalahgunaan narkoba.

Kata Kunci: Penyalahgunaan Narkoba, Logistic Regression, Support Vector Machine, Klasifikasi, Machine Learning.

PENDAHULUAN

Penyalahgunaan narkoba merupakan masalah kesehatan global yang kompleks dengan dampak kematian yang tinggi. Data United Nations Office on Drugs and Crime menunjukkan sekitar 316 juta pengguna narkoba pada 2025

dengan kesenjangan pengobatan lebih dari 80% di banyak negara. Sementara itu, World Health Organization mencatat lebih dari 3 juta kematian setiap tahun akibat alkohol dan narkoba. Kondisi ini tidak hanya merugikan secara ekonomi, tetapi juga membebani sistem kesehatan, terutama di negara berkembang.

Berbagai algoritma seperti Logistic Regression (LR) dan Support Vector Machine (SVM) telah digunakan untuk memprediksi risiko penggunaan narkoba dan menunjukkan performa yang baik. Sejumlah penelitian juga mengembangkan pendekatan fairness, explainable AI, serta penanganan data tidak seimbang. Namun, masih terdapat keterbatasan, seperti kurangnya perbandingan langsung antara LR dan SVM pada dataset UCI Drug Consumption serta minimnya integrasi interpretasi model.

Oleh karena itu, penelitian ini bertujuan membandingkan kinerja LR dan SVM dalam memprediksi risiko penggunaan narkoba berbasis klasifikasi biner. Hasilnya diharapkan dapat mendukung pengembangan alat skrining untuk intervensi dini dan sistem pendukung keputusan, sehingga berkontribusi pada kebijakan kesehatan masyarakat yang lebih efektif.

TINJAUAN PUSTAKA

Penyalahgunaan Narkoba sebagai Masalah Kesehatan Global

Penyalahgunaan narkoba merupakan salah satu isu kesehatan masyarakat global yang paling serius. Menurut *World Drug Report* 2025, pada tahun 2023 terdapat sekitar 316 juta orang yang menggunakan narkoba dalam 12 bulan terakhir, dengan peningkatan prevalensi yang melebihi pertumbuhan populasi. *World Health Organization* (WHO) melaporkan bahwa alkohol dan narkoba menyebabkan lebih dari 3 juta kematian setiap tahun, mayoritas terjadi pada kelompok laki-laki usia muda. Di Indonesia, masalah ini semakin mengkhawatirkan karena prevalensi penggunaan narkoba terus meningkat, terutama di kalangan remaja dan dewasa muda. Faktor risiko utama meliputi usia, jenis kelamin, etnis, status sosial-ekonomi, serta karakteristik kepribadian seperti *openness to experience* (Oscore) dan *sensation seeking* (SS).

Penerapan Machine Learning dalam Prediksi Resiko Penyalahgunaan Narkoba

Machine learning (ML) telah banyak digunakan untuk memprediksi risiko penyalahgunaan narkoba, khususnya *opioid use disorder* (OUD) dan konsumsi zat psikoaktif lainnya. Berbagai penelitian menunjukkan bahwa algoritma *supervised learning* mampu mengidentifikasi pola kompleks dari data demografi dan psikometrik.

Gao dkk. berhasil membangun model *logistic regression* untuk memprediksi OUD pada populasi *Medicaid* dengan akurasi yang tinggi . Faktor yang paling berpengaruh adalah dosis opioid dan determinan sosial kesehatan (SDOH). Hasan dkk. mengembangkan model dua tahap yang menggabungkan informasi demografi dan riwayat kepatuhan pengobatan untuk memprediksi penghentian pengobatan OUD.

Pada dataset UCI *Drug Consumption* yang sama dengan penelitian ini, Almahmood dkk. membandingkan 18 algoritma klasifikasi dan menemukan bahwa kombinasi model memberikan hasil optimal untuk membedakan pengguna dan bukan pengguna dari 18 jenis zat. Redhi dkk. Menerapkan *explainable AI* dengan *logistic regression* untuk memprediksi konsumsi narkoba di Bangladesh dan menekankan pentingnya interpretabilitas model.

Logistic Regression (LR)

Logistic Regression merupakan algoritma klasifikasi yang paling sering digunakan dalam prediksi risiko narkoba karena kemampuannya memberikan interpretasi yang jelas melalui *odds ratio*.

Model ini menghitung probabilitas melalui fungsi logis. Keunggulan LR adalah interpretabilitas tinggi dan kemampuan menangani data *imbalance* dengan teknik regularisasi.

Support Vector Machine (SVM)

Support Vector Machine bekerja dengan mencari hyperplane optimal untuk memisahkan kelas dengan margin maksimal. Algoritma ini sangat efektif pada data berdimensi tinggi dan imbalance.

Dalam domain prediksi narkoba, SVM sering memberikan akurasi dan *F1-score* yang lebih tinggi dibandingkan LR, meskipun interpretabilitasnya lebih rendah.

Dataset UCI Drug Consumption (Quantifed)

Penelitian ini menggunakan UCI Drug Consumption Dataset (2016) yang berisi 1.885 sampel dan 32 atribut. Fitur utama meliputi data demografi (usia, jenis kelamin, negara, etnis, pendidikan) dan karakteristik kepribadian Five Factor Model (Nscore, Escore, Oscore, Ascore, Cscore) serta impulsivity dan sensation seeking. Target variabel adalah konsumsi 18 jenis zat yang dikonversi menjadi klasifikasi biner (user/non-user).

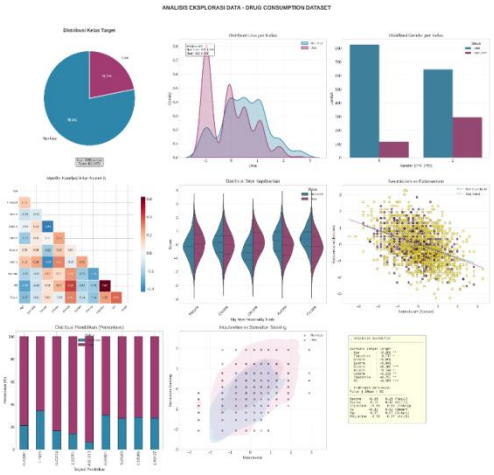
Dataset ini telah banyak digunakan dalam penelitian serupa dan sangat cocok untuk tugas multi-label classification.

METODE

Penelitian ini merupakan penelitian komputasional dengan pendekatan kuantitatif, menggunakan metode machine learning untuk klasifikasi data imbalance pada prediksi risiko penggunaan narkoba. Desain penelitian mengadopsi kerangka Cross-Industry Standard Process for Data Mining (CRISP-DM), yang terdiri dari tahap pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment. Pendekatan ini dipilih karena fleksibilitasnya dalam menangani data imbalance, sebagaimana direkomendasikan dalam review predictive modeling pada domain imbalance. Selain itu, penelitian ini bersifat eksperimental dengan fokus pada perbandingan performa algoritma, mirip dengan studi Hasan et al. yang menggunakan machine learning untuk prediksi discontinuation treatment OUD. Hasil prediksi dimaksudkan sebagai decision support untuk screening risiko tinggi dan early intervention dalam konteks global krisis opioid.

HASIL DAN PEMBAHASAN

Hasil Preprocessing dan Analisis Eksplorasi Data (EDA)



Gambar 1 Analisis Eksplorasi Data Komprehensif

Dataset Drug Consumption (UCI) diproses menjadi 1.885 sampel dengan 12 fitur setelah penghapusan kolom ID dan encoding kategorikal pada variabel Gender, Country, dan Ethnicity. Variabel target dikonversi ke format biner (0 = Non-User/CL0; 1 = User/CL1–CL6). Distribusi kelas menunjukkan ketidakseimbangan signifikan dengan proporsi Non-User 21,9% (413 sampel) dan User 78,1% (1.472 sampel). Pembagian data dilakukan secara stratified dengan rasio 80:20 untuk pelatihan (1.508 sampel) dan pengujian (377 sampel). Analisis eksplorasi data mengonfirmasi bahwa fitur kepribadian

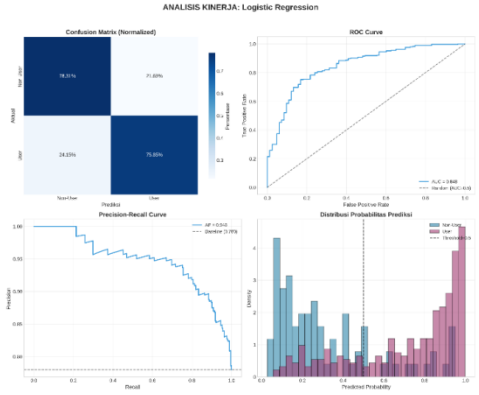
Journal of Science and Social Research ISSN 2615 – 4307 (Print)
April 2026, IX (2): 2062 – 2067 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

Sensation Seeking (SS) dan Openness to Experience (Oscore) memiliki korelasi positif terkuat terhadap target (r ≈ +0,36 dan r ≈ +0,31), dengan effect size besar pada SS (+0,92) dan Impulsive (+0,63), menjadikan kedua fitur ini sebagai prediktor utama.

Hasil Pemilihan Fitur dengan Recursive Feature Elimination (RFE)

Proses RFE dengan estimator Logistic Regression menghasilkan 8 fitur terpilih yang konsisten pada kedua model, yaitu Age, Gender, Country, Ethnicity, Escore, Oscore, Cscore, dan SS. Fitur Nscore, Ascore, dan Impulsive tidak terpilih (ranking >1). Pemilihan ini mereduksi dimensi dari 12 menjadi 8 fitur, meningkatkan efisiensi komputasi sekaligus meminimalkan risiko multikolinearitas antar fitur kepribadian.

Hasil Evaluasi Model Logistic Regression

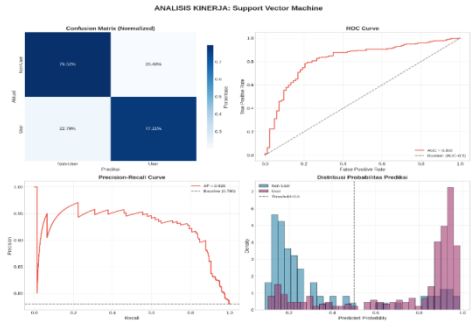


Gambar 2 Analisis Kinerja Logistic Regression

Model *Logistic Regression* dengan parameter optimal (C=10, *penalty* = L1) menunjukkan kinerja yang baik dengan nilai *cross-validation* F1-score sebesar 0,8328. Pada data uji, model mencapai *accuracy* 0,7639, *precision* 0,9253, *recall* 0,7585, dan F1-score 0,8336 (95% CI: 0,800–0,866). Nilai AUC-ROC sebesar 0,8478 (95% CI: 0,795–0,892) mengindikasikan kemampuan diskriminasi yang tinggi. Selain itu, MCC sebesar 0,4673 dan G-Mean 0,7707 menunjukkan performa yang stabil pada data tidak seimbang.

Analisis lebih lanjut melalui *confusion matrix* ter-normalisasi menunjukkan *true negative* sebesar 78,31% dan *true positive* 75,85%. Kurva ROC (AUC = 0,848) dan *precision-recall* (AP = 0,948) memperkuat temuan bahwa model memiliki *precision* tinggi dengan *false positive* yang rendah, meskipun nilai *recall* masih perlu ditingkatkan untuk meminimalkan *false negative* dalam konteks deteksi dini.

Hasil Evaluasi Model Support Vector Machine



Gambar 3 Analisis Kinerja Support Vector Machien

Model *Support Vector Machine* (SVM) dengan parameter optimal (C=1, *gamma* = *scale*) menunjukkan kinerja yang kompetitif dengan nilai *cross-validation* F1-score sebesar 0,8272. Pada data uji, model mencapai *accuracy* 0,7772, *precision* 0,9303, *recall* 0,7721, dan F1-score 0,8439 (95% CI: 0,808–0,877). Nilai AUC-ROC sebesar 0,8204 (95% CI: 0,767–0,872) mengindikasikan kemampuan diskriminasi yang baik. Selain itu, MCC sebesar 0,4919 dan G-Mean 0,7836 menunjukkan performa yang stabil pada data tidak seimbang.

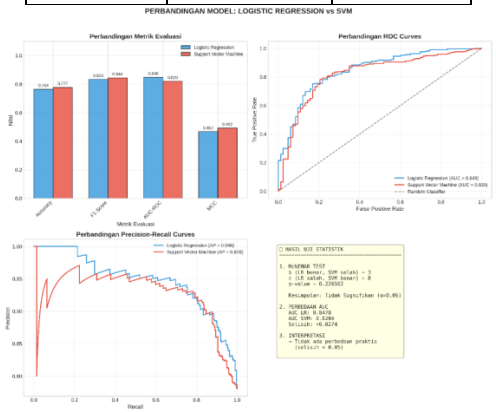
Secara komparatif, nilai *accuracy* dan F1-score yang lebih tinggi dibandingkan model sebelumnya menunjukkan bahwa SVM memiliki kinerja klasifikasi yang lebih baik secara keseluruhan. *Precision* yang tinggi (0,9303) menandakan rendahnya *false positive*, sementara *recall* 0,7721 menunjukkan kemampuan deteksi yang

cukup baik terhadap kelas positif. Analisis *confusion matrix* ter-normalisasi memperlihatkan *true negative* sebesar 79,52% dan *true positive* 77,21%, dengan *false negative* yang lebih rendah. Kurva ROC (AUC = 0,820) dan *precision-recall* (AP = 0,928) mengonfirmasi kemampuan diskriminasi yang andal, meskipun distribusi probabilitas menunjukkan sedikit *overlap* yang lebih besar dibandingkan model *Logistic Regression*.

Perbandingan Kinerja Dua Model

Tabel 1 Perbandingan Kinerja

Metrik	Logistic Regression	Support Vector Machine
Accuracy	0.7639	0.7772
F1-Score	0.8336	0.8439
AUC-ROC	0.8478	0.8204
MCC	0.4673	0.4919



Gambar 4 Perbandingan Kinerja Model LR dan SVM

Gambar 4 menunjukkan perbandingan: Bar chart metrik memperlihatkan SVM unggul pada Accuracy, F1-Score, dan MCC, sementara LR lebih baik pada *AUC-ROC*. Kurva ROC gabungan menunjukkan LR sedikit lebih tinggi di sebagian besar *false positive* rate. *Precision-Recall* curves menunjukkan LR AP=0.948 > SVM AP=0.928. Uji McNemar (b=3, c=8, *p-value*=0.226562) menunjukkan tidak ada perbedaan signifikan (p>0.05). Selisih AUC +0.0274 (<0.05) menandakan tidak ada perbedaan praktis. SVM lebih unggul tipis pada metrik keseluruhan, tapi LR lebih baik pada diskriminasi probabilitas.

Pembahasan

Perbandingan kinerja menunjukkan bahwa model SVM unggul tipis dibandingkan *Logistic Regression* pada F1-score (0,8439 vs 0,8336), *accuracy* (0,7772 vs 0,7639), dan MCC (0,4919 vs 0,4673), sementara *Logistic Regression* memiliki keunggulan pada AUC-ROC (0,8478 vs 0,8204). Namun, perbedaan tersebut tidak signifikan baik secara statistik (uji McNemar, *p*-value = 0,226562) maupun secara praktis (selisih AUC < 0,05). Kedua model menunjukkan *precision* tinggi (>92%), yang mengindikasikan kemampuan baik dalam meminimalkan *false positive*, sehingga relevan untuk aplikasi skrining.

Variabel kepribadian, khususnya *Openness to Experience* (Oscore) dan *Sensation Seeking* (SS), konsisten menjadi prediktor dominan, sejalan dengan teori psikologi yang mengaitkan keterbukaan tinggi dan kecenderungan mencari sensasi dengan perilaku berisiko seperti penggunaan *cannabis*. Penerapan teknik SMOTE dalam pipeline terbukti efektif dalam menangani ketidakseimbangan data, yang tercermin dari nilai G-Mean yang relatif tinggi (>0,77) pada kedua model.

Dari sisi implementasi, *Logistic Regression* menawarkan keunggulan interpretabilitas melalui koefisien dan *odds ratio*, sehingga lebih sesuai untuk konteks klinis dan edukatif. Sebaliknya, SVM memberikan performa prediktif sedikit lebih baik namun bersifat *black-box* dan kurang transparan.

Keterbatasan penelitian ini meliputi penggunaan dataset UCI yang bersifat umum dan tidak spesifik pada konteks Indonesia, pendekatan klasifikasi biner yang sederhana, serta ketergantungan pada data *self-reported* yang berpotensi menimbulkan *recall bias*.

SIMPULAN

Penelitian ini berhasil membandingkan kinerja algoritma Logistic Regression dan Support Vector Machine dalam memprediksi risiko penggunaan cannabis menggunakan dataset Drug Consumption (UCI Machine Learning Repository) dengan total 1.885 sampel. Setelah preprocessing, distribusi kelas menunjukkan ketidakseimbangan signifikan (NonUser 21.9%, User 78.1%), yang ditangani efektif melalui SMOTE. Analisis eksplorasi data mengonfirmasi bahwa fitur kepribadian seperti Oscore (Openness to Experience) dan SS (Sensation Seeking) merupakan prediktor utama, dengan korelasi positif moderat dan effect size besar pada SS (+0.92). Pemilihan fitur menggunakan RFE menghasilkan 8 fitur terbaik yang sama pada kedua model, yaitu Age, Gender, Country, Ethnicity, Escore, Oscore, Cscore, dan SS. Hasil evaluasi menunjukkan kedua model memiliki performa yang baik dengan *F1-Score* di atas 0.83 dan AUC-ROC di atas 0.82; Support Vector Machine unggul tipis pada Accuracy (0.7772), *F1-Score* (0.8439), dan MCC (0.4919), sedangkan Logistic Regression lebih baik pada AUC-ROC (0.8478) serta memberikan interpretabilitas tinggi melalui odds ratio (Oscore 1.9694 atau peningkatan risiko 96.9%, SS 1.7979 atau 79.8%). Perbandingan kinerja menunjukkan tidak ada perbedaan signifikan secara statistik (uji McNemar p-value = 0.226562 > 0.05) maupun praktis (selisih AUC = 0.0274 < 0.05). Learning curves mengindikasikan kedua model stabil tanpa overfitting berarti (gap akhir LR -0.005 dan SVM +0.021). Secara keseluruhan, pendekatan machine learning terbukti efektif untuk skrining dini risiko penggunaan cannabis dengan fokus pada faktor kepribadian, meskipun masih terdapat keterbatasan pada representasi data UCI yang bersifat umum.

DAFTAR PUSTAKA

Eccles, R. G., & Klimenko, S. (2019). The investor revolution. Harvard Business Review, 97(3), 106–116.

Elkington, J. (1997). Cannibals with forks: The triple bottom line of 21st century business. Capstone Publishing.

Global Reporting Initiative. (2021). GRI sustainability reporting standards. https://www.globalreporting.or

Kaplan, R. S., & Norton, D. P. (2001). The strategy-focused organization: How balanced scorecard companies thrive in the new business environment. Harvard Business School Press. Laudon, K. C., & Laudon, J. P. (2020). Management information systems: Managing the digital firm (16th ed.). Pearson.

Porter, M. E., & Kramer, M. R. (2011). Creating shared value. Harvard Business Review, 89(1–2), 62–77.

Searcy, C., & Elkhawas, D. (2012). Corporate sustainability ratings: An investigation into how corporations use the Dow Jones Sustainability Index. Journal of Cleaner Production, 35, 79–92. https://doi.org/10.1016/j.jclepro.2012.05.022

United Nations. (2015). Transforming our world: The 2030 agenda for sustainable development. https://sdgs.un.org

World Economic Forum. (2020). Measuring stakeholder capitalism: Towards common metrics and consistent reporting of sustainable value creation. https://www.weforum.org

Zhou, X., Simnett, R., & Green, W. (2017). Does integrated reporting matter to the capital market? Abacus, 53(1), 94–132. https://doi.org/10.1111/abac.12104