JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 1, No. 11, Tahun 2024
ISSN 3025-0919 (media online)
Hal 1131-1138

Literature Review: Perbandingan Metode Klasifikasi Dalam *Data Mining*

Dandi Pangestu[1], Nabilah Nur Zakiyyah[2], Nurul Fauziah[3*], Zulaizah Rahayu[4], Ines Heidiani Ikasari[5]

[1]Fakultas Ilmu Komputer, Program Studi Informatika, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
Email: [1]dandipngst@gmail.com, [2]nabilahnurzkiyyah@gmail.com, [3*]nurulfauzhh@gmail.com, [4]zr.ayyu@gmail.com, [5]dosen01374@unpam.ac.id
(* : coressponding author)

Abstrak − Pada penelitian ini penulis menganalisis 10 jurnal terkait implementasi *Data Mining* pada berbagai aspek. *Data mining* sendiri merupakan teknik penting dalam menghasilkan informasi berharga dari data yang besar. *Data mining* banyak diimplementasikan dalam banyak aspek kehidupan seperti aspek kesehatan, akademik atau pendidikan, dan ekonomi. Penelitian dilakukan dengan membandingkan berbagai metode klasifikasi data mining dalam memprediksi kinerja akademik mahasiswa. Metode yang dibahas meliputi *Naive Bayes*, *K-Nearest Neighbors* (KNN), *Support Vector Machines* (SVM), dan *Decision Tree*. Tujuan penelitian ini adalah untuk melakukan kajian pustaka dengan mengidentifikasi subjek atau dataset yang digunakan, metode analisis data yang diterapkan, tahun penelitian, dan tingkat akurasi yang dicapai dalam jurnal-jurnal yang dianalisis. Dengan menggunakan metode yang disajikan, penulis berusaha untuk menyimpulkan hasil penelitian guna mencari tingkat akurasi tertinggi dan mengidentifikasi peluang penelitian lanjutan.

Kata Kunci: Data Mining; Metode Klasifikasi; Evaluasi

Abstract *− In this study, the authors analyzed 10 journals related to the implementation of Data Mining. Data mining is an important technique in extracting valuable information from large data. Data mining is widely implemented in many aspects of life such as health, academic or educational, and economic aspects. Research was conducted by comparing various data mining classification methods in predicting student academic performance. The methods discussed include Naive Bayes, K-Nearest Neighbors (K-NN), Support Vector Machines (SVM), and Decision Tree. The purpose of this research is to conduct a literature review by identifying the subjects or datasets used, data analysis methods applied, years of research, and the level of accuracy achieved in the journals analyzed. Using the methods presented, the authors attempted to summarize the research results to find the highest level of accuracy and identify further research opportunities.*

Keywords*: Data Mining; Classification Methods; Evaluation*

1. PENDAHULUAN

Data Mining atau yang dapat disebut juga dengan penambangan data merupakan sebuat proses untuk menemukan pola dan aturan dalam kumpulan data besar yang sebelumnya tidak diketahui melalui dengan teknik analisis data tingkat lanjut (Annisa, 2019). Ada beberapa tahapan yang harus dilakukan pada proses data mining yaitu menyeleksi data, kemudian dilakukan processing agar kualitas data menjadi lebih baik, selanjutnya dilakukan transformasi, dan interpretasi, serta yang terakhir adalah tahap evaluasi, agar keluaran yang dihasilkan menjadi sebuah pengetahuan baru yang kemudian dapat memberikan dampak lebih baik. *Data mining* juga memiliki peran penting seperti klasifikasi dan prediksi (De Wibowo Muhammad Sidik et al., 2020).

Dalam beberapa tahun terakhir, banyak metode data mining yang dikembangkan dan direkomendasikan untuk digunakan dalam proses klasifikasi dan prediksi. Metode-metode ini mencakup *naive bayes*, *decision tree*, *neural network*, *k-nearest neighbour*, dan *support vector machine*. Kelima metode ini pada dasarnya memiliki keunggulan masing-masing. Metode *naive bayes* dapat diterapkan pada berbagai jenis data (seperti kuantitatif dan kualitatif) dan data pelatihan yang memerlukan jumlah data yang relatif kecil, sedangkan pohon keputusan memiliki sensitivitas yang sangat tinggi dan dapat digunakan untuk menggabungkan data dengan beberapa dimensi. Neural network merupakan metode dengan kurva pembelajaran yang relatif kecil karena dapat memproses data secara bolak-balik (De Wibowo Muhammad Sidik et al., 2020).

Selanjutnya metode K-NN merupakan salah satu algoritma sederhana untuk memecahkan

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 1, No. 11, Tahun 2024
ISSN 3025-0919 (media online)
Hal 1131-1138

suatu masalah klasifikasi, algoritma K-NN dapat menghasilkan hasil yang kompetitif dan signifikan (Analisis Dan Penerapan et al., 2019), dan terakhir *support vector machine*, merupakan suatu metode yang dapat menghasilkan hasil yang cukup maksimal meskipun hanya memiliki sedikit data training, karena untuk melatih metode tersebut hanya diperlukan data yang relatif sedikit (Pangestu et al., 2023).

Dalam melakukan perbandingan dan evaluasi kinerja pengklasifikasi, terdapat beberapa metrik yang dapat digunakan untuk mengukur sejauh mana model atau sistem tersebut efektif dan kredibel. Metrik-metrik ini meliputi akurasi, presisi, recall, dan F1-score. Akurasi dapat mengukur seberapa tepat model dalam mengklasifikasikan seluruh instance dengan benar.

Sementara itu, presisi mengukur proporsi instance positif yang secara benar diklasifikasikan oleh model dari keseluruhan instance yang diprediksi positif. Recall, di sisi lain, mengukur proporsi instance positif yang benar diklasifikasikan oleh model dari keseluruhan instance yang seharusnya positif. F1-score adalah harmonic mean dari presisi dan recall, memberikan gambaran komprehensif tentang kinerja model dalam mengklasifikasikan instance positif. Dengan menggunakan kombinasi metrik-metrik ini, peneliti dapat mendapatkan pemahaman yang lebih mendalam tentang kekuatan dan kelemahan suatu model pengklasifikasi dalam menangani tugas-tugas klasifikasi yang kompleks.

Penelitian ini bertujuan untuk membandingkan dan mengevaluasi kinerja beberapa metode klasifikasi *Data Mining* yang populer, yaitu *Naive Bayes, Decision Tree, Neural Network, K-Nearest Neighbor,* Dan *Support Vector Machine.*

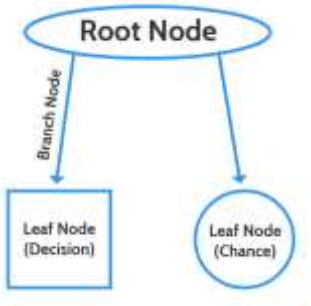
2. TINJAUAN PUSTAKA

2.1 Metode Klasifikasi

Metode Klasifikasi dalam *Data Mining* merupakan suatu teknik yang digunakan untuk mengelompokkan data ke dalam kelas atau kategoru tertentu berdasarkan karakteristik atau atribut tertentu (Gede et al., 2020). Dalam metode ini ada beberapa algoritma yang bisa digunakan seperti *Decision Trees, Naive Bayes, support vector machines (SVM), K-Nearest Neighbors (KNN), Neural Network* dan algoritma klasifikasi lainnya.

a. *Decision Tree*

Salah satu algoritma dalam klasifikasi yang digunakan untuk mengambil keputusan dengan menerjemahkan data menjadi struktur pohon yang memiliki cabang-cabang dan daun-daun sebagai hasil klasifikasi. Visualisasi ini dapat memudahkan untuk memahami hubungan antar data. *Decision Tree*, atau Pohon Keputusan, tersusun atas node-node yang saling terhubung dan membentuk struktur seperti pohon dengan akar. Akar merupakan simpul utama, dan node-node lainnya terhubung ke akar melalui cabang-cabang. Node dengan cabang keluar (memiliki anak) disebut node internal atau node tes, sedangkan node yang tidak memiliki cabang keluar (tidak memiliki anak) disebut node daun. Dalam pohon keputusan, setiap simpul internal membagi ruang menjadi dua atau lebih sub ruang sesuai dengan fungsi diskrit tertentu dari atribut nilai (Purwati et al., 2020).



Gambar 1. Konsep *Decision Trees* (*Source:https://venngage.com/*)

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 1, No. 11, Tahun 2024
ISSN 3025-0919 (media online)
Hal 1131-1138

b. *Naive Bayes*

Merupakan teknik prediksi berbasis probabilitas sederhana yang berdasarkan pada penerapan aturan bayes dengan asumsi ketidaktergantungan yang kuat atau bersifat bebas (independence) (Subarkah et al., 2020). Digunakan karena mudah diimplementasikan dan memiliki hasil yang baik saat diterapkan pada banyak kasus. Sedangkan kelemahan dari metode ini yaitu adanya asumsi atau dengan kata lain kondisi kelas saling bebas, sehingga kurang akurat (Anjelika, 2021).

Rumus teorama Bayes:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

c. *Support Vector Machines* (SVM)

SVM adalah algoritma yang bekerja menggunakan pemetaan nonlinier untuk mengubah data pelatihan asli ke dimensi yang lebih tinggi. Dimensi baru ini akan mencari hyperplane untuk memisahkan secara linier dengan pemetaan nonlinier ke dimensi yang cukup tinggi. Persoalan nonlinear merupakan persoalan dengan data yang tidak dapat dipisahkan secara linear yaitu tidak ada sebuah garis yang dapat dibuat sebagai pemisah antar kelas data (Nikmatun & Waspada, 2019). Fungsi pemisah yang terbaik yaitu dapat mengoptimalkan nilai margin yang merupakan separating hyperplane pada setiap kelas dan posisi ini dapat tercapai apabila garis pemisah tersebut posisinya tepat berada di tengah-tengah, kemudian membagi antar kelas negatif dan kelas positif (Pangestu et al., 2023).

d. *K-Nearest Neighbors* (KNN)

KNN adalah algoritma klasifikasi non-parametrik dasar tetapi efektif. Disamping itu, KNN memiliki kelemahan yang signifikan. Dalam banyak aplikasi, seperti penambangan web dinamis untuk repository yang luas, efisiesinya yang rendah mencegahnya digunakan karena merupakan metode pembelajaran yang malas. Misalnya mengindeks instance pelatihan sebagai pengklasifikasian KNN mengharuskan penyimpanan seluruh set pelatihan saat ini tdak pada redudansi set pelatihan untuk mengurangi kesulitan ini dapat sangat meminimalkan komputasi yang diperlukan pada waktu kueri (Annisa, 2019).

e. *Neural Network*

Neural network (jaringan saraf tiruan) adalah model non-linear yang cukup rumit dibangun dari komponen yang secara individu berperilaku mirip seperti model regresi. *Neural network* dapat direpresentasikan suatu grafik, dan beberapa sub-grafik tampaknya ada integritas yang sama dengan gerbang logika. Struktur dari jaringan neuron atau saraf secara terperinci dirancang terlebih dahulu (Chatrina Siregar et al., 2020).

3. METODE

3.1 Desain penelitian

Jurnal penelitian ini memiliki tujuan untuk menganalisis isi tinjauan pustaka sebelumnya, yang berfokus pada temuan berbagai penelitian yang telah dipublikasikan dari berbagai jurnal nasional.

3.2 Sumber data

Data yang dikumpulkan untuk analisis ini berasal dari berbagai jurnal nasional yang membahas tentang metode-metode klasifikasi yang digunakan pada data mining. Sumber jurnal tersebut termasuk Google Scholar (*https://scholar.google.com/*).

Dalam penelitian ini, penulis mengumpulkan semua isi tinjauan pustaka yang membahas mengenai metode klasifikasi dalam *Data Mining*. Total 10 jurnal yang dipilih yang telah dipublikasikan secara online dan mengulas metode klasifikasi ini. Semua tinjauan literatur ini kemudian dianalisis untuk menentukan metode apa yang paling efektif untuk digunakan dalam

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 1, No. 11, Tahun 2024
ISSN 3025-0919 (media online)
Hal 1131-1138

pengklasifikasian *data mining.*

Tabel 1. *Aspects and categories used for content*

Aspects	***Categories***
Kind of Research	A.1 *Experiment* A.2 *Literature review*
Year	B.1 2019 B.2 2020 B.3 2021 B.4 2023
Dataset	C.1 *Public* C.2 *Private*
Methods	D.1 *Decision Tree* D.2 NBC D.3 SVM D.4 KNN D.5 *Neural Network*
Accuracy	E.1 *<40% Low* E.2 *40% – 80% Middle E.3 >80% Hight*
Scope	F.1 *Health* F.2 *Academic* F.3 *General* F.4 *Economics*

3.3. Instrumen Penelitian

Penelitian ini merupakan sebuah studi literatur yang mencakup beberapa jurnal penelitian terkait metode klasifikasi dalam data mining. Tinjauan ini bertujuan untuk melihat upaya penelitian terbaru yang telah mengimplementasikan metode klasifikasi. Dalam penelitian ini, kami mengumpulkan literatur dari berbagai sumber yang mencakup berbagai metode klasifikasi yang digunakan dalam jurnal penelitian. Proses pengumpulan data ini penting untuk menemukan dan mendapatkan referensi kajian yang relevan berdasarkan penelitian sebelumnya.

Tabel 1 dalam jurnal penelitian ini memuat beberapa aspek utama yang diperoleh dari tinjauan pustaka beberapa jurnal nasional. Aspek-aspek tersebut meliputi:

a. Jenis penelitian

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 1, No. 11, Tahun 2024
ISSN 3025-0919 (media online)
Hal 1131-1138

b. Tahun penelitian

c. Dataset yang digunakan

d. Metode penelitian yang diterapkan

e. Akurasi hasil penelitian

f. Bidang penelitian yang terkait

Tabel ini memberikan informasi penting mengenai karakteristik penelitian mengenai metode klasifikasi dalam data mining.

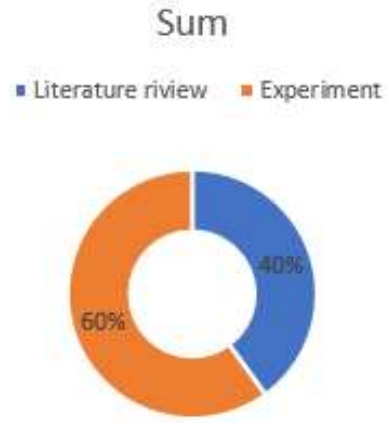
4. ANALISA DAN PEMBAHASAN

Berdasarkan grafik yang tercantum dalam tabel 1, jumlah publikasi artikel menggambarkan tingkat akurasi penelitian yang dilakukan dalam periode tertentu. Terlihat bahwa beberapa metode klasifikasi mencapai tingkat akurasi rata-rata tertinggi hingga 90%, sedangkan ada juga yang memiliki tingkat akurasi rata-rata sebesar 60%.

Selain itu, setiap tahunnya terdapat publikasi yang menggunakan metode yang berbeda-beda. Hal ini menunjukkan bahwa data mining dapat diimplementasikan di berbagai bidang, tidak hanya terbatas pada bidang akademik saja. Perkembangan penelitian ini menunjukkan bahwa ada beragam aplikasi data mining dalam konteks yang lebih luas.

4.1 Jenis Penelitian

Beberapa peneliti memanfaatkan pendekatan penelitian experiment dan literature review. Dalam jenis penelitian experiment ini, hampir semua peneliti menggunakan database baik yang bersifat publik maupun tidak publik. Di samping itu, sebagian peneliti juga memiliki minat yang lebih kuat dalam mempelajari data mining secara teoritis, sehingga penulis juga melakukan tinjauan pustaka terhadap beberapa penelitian dari berbagai sumber.

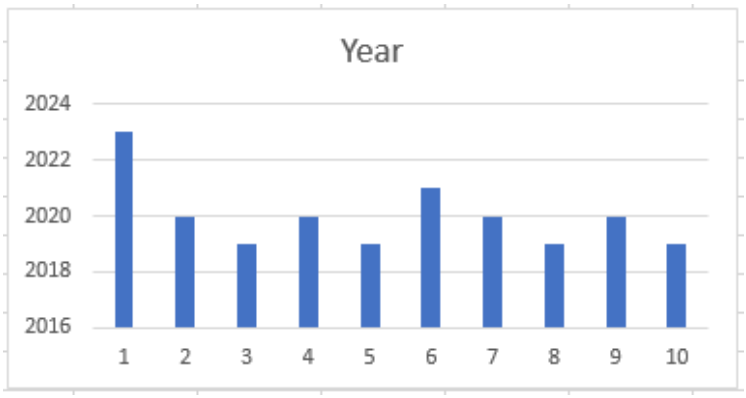


Gambar 2. Diagram Jenis Penelitian

Pada Gambar 3, ada 2 tipe penelitian yang digunakan yaitu literature review dan experiment. Penelitian ini meneliti 10 jurnal dengan jumlah terbesar yaitu 60% pada eksperimen dan 40% literature review.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 1, No. 11, Tahun 2024
ISSN 3025-0919 (media online)
Hal 1131-1138

4.2 Tahun Penelitian

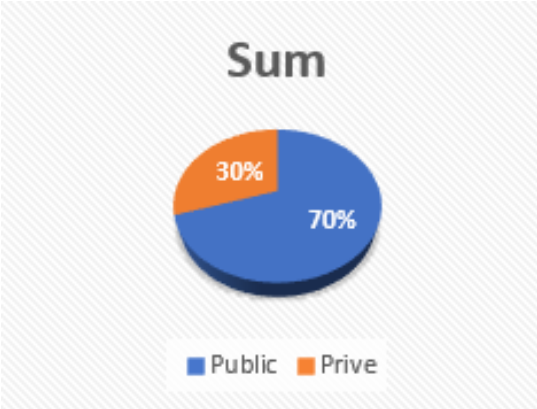


Gambar 3. Tahun Penelitian

Berdasarkan penelitian pada Gambar 3, terdapat beberapa tahun yang menjadi fokus penelitian, termasuk tahun 2019, 2020, 2021, dan 2023. Dari sekian tahun tersebut, penelitian paling banyak dilakukan pada tahun 2019 dan 2020 dengan jumlah penelitian masing-masing sebanyak 4 penelitian. Sementara itu, tahun 2021 dan 2023 terdapat satu penelitian. Meskipun begitu, penelitian pada tahun 2022 masih tergolong sedikit. Oleh karena itu, ini memberikan peluang bagi penelitian lebih lanjut pada tahun 2022.

4.3 Datasets

Penelitian ini menggunakan kombinasi dataset publik dan dataset privat Gambar 5. Dataset publik merujuk pada himpunan data yang dapat dengan mudah ditemukan dan diakses. Sementara itu, dataset pribadi merujuk pada himpunan data yang sulit ditemukan dan membutuhkan persetujuan tertentu untuk mendapatkannya. Dalam penelitian ini, terdapat dua dataset pribadi dan delapan dataset publik yang digunakan.

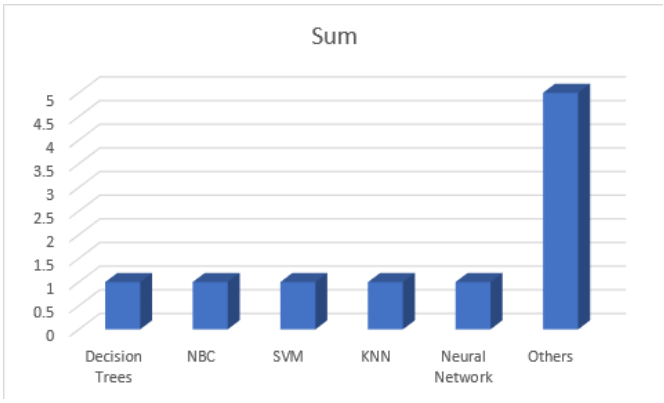


Gambar 4. Dataset yang digunakan

4.4 Metode Penelitian

Pada penelitian ini, penulis menganalisis metode yang digunakan meliputi *Decision Tree*, *Naïve Bayes Classifier* (NBC), *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), dan *Neural Network*. Dari 10 jurnal yang penulis analisis, 5 diantaranya menggunakan lebih dari satu metode klasifikasi yang berarti satu penelitian menggunakan 2 atau lebih metode klasifikasi sebagai komparasi. Sedangkan, 5 penelitian lainnya menggunakan metode-metode klasifikasi yang disebutkan sebelumnya.

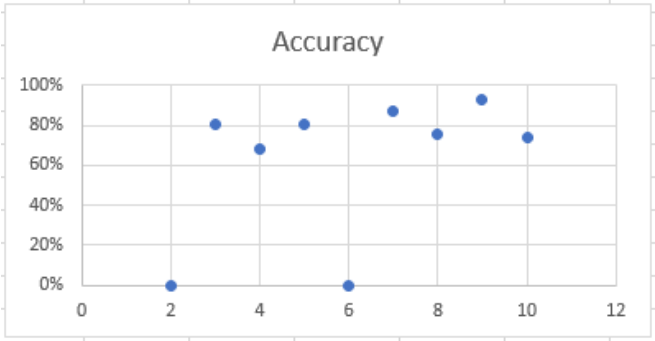
JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 1, No. 11, Tahun 2024
ISSN 3025-0919 (media online)
Hal 1131-1138



Gambar 5. Metode Penelitian

4.5 Akurasi Penelitian

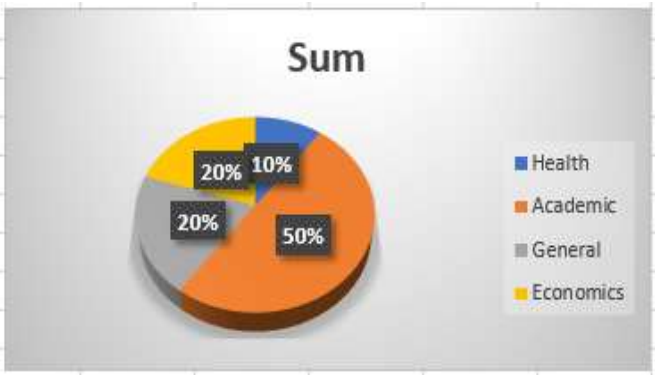
Rata-rata tingkat akurasi pada Gambar 5 melampaui 60%. Beberapa penelitian yang telah dianalisis tidak memberikan informasi terperinci tentang tingkat akurasi yang dicapai, melainkan hanya menjelaskan proses pengumpulan data atau perbandingan metode. Dalam 10 jurnal yang dianalisis, tujuh di antaranya melaporkan tingkat akurasi yang dicapai. Salah satunya mencapai 92.83% (Chatrina Siregar et al., 2020), sedangkan jurnal lainnya mencapai 68% (Anjelika, 2021).



Gambar 6. Tingakatan Akurasi

4.6 Bidang Penelitian

Dalam penelitian ini, penulis menganalisis empat bidang penelitian yang melibatkan aspek kesehatan, akademik, ekonomi dan umum. Berdasarkan Gambar 6, penelitian tersebut menunjukkan bahwa sektor akademik mencakup 50% dari bidang penelitian yang dianalisis, sedangkan sektor ekonomi dan umum mencakup 20%, dan sector kesehatan mencakup 10%. Hal ini menunjukkan bahwa penulis memberikan perhatian lebih pada bidang akademik.



Gambar 7. Diagram Bidang Penelitian

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 1, No. 11, Tahun 2024
ISSN 3025-0919 (media online)
Hal 1131-1138

5. KESIMPULAN

Kesimpulan pada literture review ini yaitu, dalam perbandingan beberapa metode klasifikasi, setiap metode menghasilkan hasil akurasi yang berbeda-beda. Akurasi tertinggi pada 10 jurnal yang penulis analisis dihasilkan jurnal yang menggunakan metode klasifikasi neural network dengan presentase mencapai 92.83%. Tetapi tidak menutup kemungkin jika metode yang lain juga memiliki presentase yang sama atau lebih besar, dikarenakan jurnal yang penulis analisis hanya sedikit.Hasil penelitian juga menunjukkan bahwa tidak ada metode yang terbaik untuk semua dataset, karena setiap dataset memiliki karakteristik yang berbeda-beda sehingga perlu diperhatikan pemilihan metode untuk setiap dataset.

REFERENCES

Analisis Dan Penerapan, ., Handayanto, A., Latifa, K., Saputro, N. D., & Waliyansyah, R. R. (2019). *Analisis dan Penerapan Algoritma Support Vector Machine (SVM) dalam Data Mining untuk Menunjang Strategi Promosi (Analysis and Application of Algorithm Support Vector Machine (SVM) in Data Mining to Support Promotional Strategies)* (Vol. 7, Issue 2).

Anjelika, Y. (n.d.). *LITERATUR REVIEW : PREDIKSI KELULUSAN MAHASISWA DENGAN MENGGUNAKAN ILMU DATA MINING*. https://www.researchgate.net/publication/351342175

Annisa, R. (2019). ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK PREDIKSI PENDERITA PENYAKIT JANTUNG. *Jurnal Teknik Informatika Kaputama (JTIK)*, *3*(1).

Chatrina Siregar, N., Ruli, R., Siregar, A., Yoga, ; M, & Sudirman, D. (2020). Implementasi Metode Naive Bayes Classifier (NBC) Pada Komentar Warga Sekolah Mengenai Pelaksanaan Pembelajaran Jarak Jauh (PJJ). In *Jurnal Teknologia Aliansi Perguruan Tinggi (APERTI) BUMN* (Vol. 3, Issue 1).

De Wibowo Muhammad Sidik, A., Himawan Kusumah, I., Suryana, A., Artiyasa, M., & Pradiftha Junfithrana, A. (2020). *Gambaran Umum Metode Klasifikasi Data Mining*. *2*(2), 34–38.

Gede, I., Sudipa, I., Profile, S., & Darmawiguna, M. (n.d.). *BUKU AJAR DATA MINING*. https://www.researchgate.net/publication/377415198

Nikmatun, I. A., & Waspada, I. (2019). IMPLEMENTASI DATA MINING UNTUK KLASIFIKASI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR. *Jurnal SIMETRIS*, *10*(2).

Pangestu, P., Novita, R., Informasi, S., & Sultan Syarif Kasim Riau, U. (n.d.). *Systematic Literature Review: Perbandingan Algoritma Klasifikasi*. *8*(2), 2023.

Purwati, N., Nurlistiani, R., & Devinsen, O. (2020). DATA MINING DENGAN ALGORITMA NEURAL NETWORK DAN VISUALISASI DATA UNTUK PREDIKSI KELULUSAN MAHASISWA. *Jurnal Informatika*, *20*(2), 156–163. https://doi.org/10.30873/ji.v20i2.2273

Subarkah, P., Pambudi, E. P., & Hidayah, S. O. N. (2020). Perbandingan Metode Klasifikasi Data Mining untuk Nasabah Bank Telemarketing. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *20*(1), 139–148. https://doi.org/10.30812/matrik.v20i1.826