Vol. 6, No. 1, April 2025, hal. 39-44

Jurnal Computer Science and Information Technology (CoSciTech)

p-ISSN: 2723-567X
e-ISSN: 2723-5661

http://ejurnal.umri.ac.id/index.php/coscitech/index

Analisis Data Mining Untuk Deteksi Diabetes Mellitus Menggunakan Naïve Bayes

Yusril Haza Mahendra*[1], Ririen Kusumawati[2], Imamudin[3]
Email: [1]yusrilhaza99@gmail.com, [2] ririen.kusumawati@ti.uin-malang.ac.id, [3]imamudin@ti.uin-malang.ad.id

[1] Program Studi Magister Informatika, Universitas Islam Negri Maulana Malik Ibrahim Malang
[2] Program Studi Magister Informatika, Universitas Islam Negri Maulana Malik Ibrahim Malang
[3] Program Studi Magister Informatika, Universitas Islam Negri Maulana Malik Ibrahim Malang

Diterima: 05 April 2020 | Direvisi: 05 Mei 2020 | Disetujui: 27 Mei 2020

Abstrak

Diabetes mellitus adalah penyakit kronis dengan prevalensi global yang terus meningkat. Penyakit ini ditandai oleh kadar glukosa darah yang tinggi akibat ketidakmampuan tubuh memproduksi atau menggunakan insulin secara optimal. Dampaknya yang luas terhadap individu dan masyarakat menunjukkan pentingnya deteksi dini dan pengelolaan yang tepat. Dalam era digital, penerapan *data mining* menjadi alat yang sangat penting di bidang kesehatan. Teknik ini memungkinkan analisis data skala besar untuk mengidentifikasi pola dan tren yang sulit dikenali secara manual. Dalam kaitannya dengan deteksi diabetes mellitus, *data mining* memiliki potensi signifikan untuk pengembangan model prediktif. Salah satu algoritma yang sering digunakan adalah Naïve Bayes. Penelitian ini berfokus pada analisis algoritma Naïve Bayes untuk mengklasifikasikan gejala awal diabetes mellitus, dengan tujuan memperdalam pemahaman tentang faktor risiko dan menciptakan alat deteksi dini yang efektif. Hasil penelitian menunjukkan bahwa Naïve Bayes mencapai akurasi tertinggi sebesar 78% saat menggunakan teknik imputasi nilai hilang dengan metode *mean*. Diharapkan penelitian ini dapat mendukung upaya pencegahan dan pengelolaan diabetes mellitus, serta membantu mengurangi beban bagi individu dan masyarakat secara keseluruhan.

Kata kunci: Diabetes Mellitus, Analisis Data Mining, Naïve Bayes

Data Mining Analysis for Detecting Diabetes Mellitus Using Naïve Bayes

Abstract

Diabetes mellitus is a chronic condition with a globally increasing prevalence. It is marked by elevated blood glucose levels resulting from the body's inability to produce or efficiently use insulin. The significant impact on individuals and communities highlights the necessity of early detection and effective management. In the digital age, data mining has emerged as an essential tool in healthcare, enabling large-scale analysis of health data to uncover patterns and trends that are not easily identified manually. For diabetes mellitus detection, data mining presents substantial potential for developing predictive models. Among the algorithms utilized is Naïve Bayes. This study examines the Naïve Bayes classification method for identifying early symptoms of diabetes mellitus, aiming to deepen understanding of risk factors and create effective early detection tools. The results reveal that the Naïve Bayes algorithm achieves its highest accuracy of 78% using the missing value imputation mean technique. This research is expected to support efforts in diabetes prevention and management while alleviating the burden on individuals and communities at large.

Keywords*: Diabetes Mellitus, Analisis Data Mining, Naïve Bayes*

1. PENDAHULUAN

Diabetes mellitus merupakan salah satu penyakit kronis yang prevalensinya terus meningkat secara global. Penyakit ini ditandai oleh tingginya kadar glukosa dalam darah yang disebabkan oleh ketidakmampuan tubuh untuk memproduksi atau menggunakan insulin secara efektif.[1] Dampaknya yang luas terhadap kesehatan individu dan masyarakat membuat deteksi dini serta manajemen yang tepat menjadi sangat penting.[2] Dalam era digital dan kemajuan teknologi informasi, analisis data mining telah menjadi alat yang sangat berguna dalam bidang kesehatan. Data mining memungkinkan peneliti dan praktisi kesehatan untuk mengeksplorasi dan menganalisis data kesehatan dalam skala besar untuk mengidentifikasi pola, hubungan, dan tren yang mungkin sulit dideteksi secara manual. *Diabetes mellitus* adalah salah satu penyakit kronis yang ditandai dengan kadar glukosa darah yang tinggi akibat gangguan pada produksi *insulin* atau resistensi *insulin*. Penyakit ini berkembang perlahan dan umumnya terjadi pada orang dewasa, meskipun kini kasus pada remaja dan anak-anak semakin meningkat. Faktor risiko utama meliputi *obesitas*, gaya hidup tidak aktif, serta faktor *genetik*. Diagnosis dini dan penanganan yang tepat sangat penting untuk mencegah komplikasi serius, seperti penyakit jantung, kerusakan saraf, dan kerusakan ginjal.[3] Diabetes tipe 1 adalah penyakit *autoimun* dimana sistem kekebalan tubuh secara keliru menyerang sel *beta* di pankreas, menyebabkan kerusakan permanen sehingga sel-sel tersebut tidak bisa lagi memproduksi *insulin*. Penyebabnya meliputi faktor *genetik*, lingkungan, dan gaya hidup. Diabetes tipe 2 terjadi karena resistensi *insulin*, dimana tubuh tidak merespons *insulin* dengan baik. Pankreas awalnya memproduksi lebih banyak *insulin* untuk mengatasi resistensi ini, tetapi akhirnya produksinya menurun, sehingga kadar gula darah meningkat. Faktor penyebab utamanya adalah *genetik*, kelebihan berat badan, obesitas, dan gaya hidup yang kurang aktif [4]. *Diabetes gestasional* muncul selama kehamilan akibat hormon yang mengganggu kerja *insulin*. Faktor risiko termasuk riwayat pradiabetes dan keluarga dengan diabetes. Sekitar 50% kasus diabetes gestasional dapat dikonfirmasi dari sumber yang terpercaya.

Dalam konteks deteksi diabetes mellitus, analisis data mining memiliki potensi besar untuk membantu dalam pengembangan model prediktif yang dapat mengidentifikasi individu yang berisiko tinggi untuk mengembangkan diabetes atau membantu dalam manajemen penyakit pada individu yang sudah didiagnosis.[1] Pendekatan data mining untuk deteksi diabetes mellitus melibatkan penggunaan berbagai teknik, termasuk tetapi tidak terbatas pada klasifikasi, klastering, dan asosiasi. salah satu algoritma data mining adalah naïve bayes.[5] Dengan memanfaatkan dataset yang mencakup informasi tentang faktor risiko seperti riwayat keluarga, pola makan, aktivitas fisik, dan hasil tes laboratorium, model-model prediktif dapat dikembangkan untuk mengidentifikasi individu yang rentan terhadap diabetes.[6]

Penelitian yang dilakukan oleh Madhubala *et al.* (2022) bertujuan untuk memprediksi diabetes menggunakan jaringan saraf tiruan yang ditingkatkan menggunakan multilayer perceptron, tujuan dari penelitian ini adalah untuk membangun model jaringan saraf yang dalam menggunakan dataset pima indian diabetes untuk klasifikasi diabetes.[7] Penelitian yang dilakukan oleh Shafaeizadeh *et al*. (2020) bertujuan untuk mengevaluasi apakah diabetes mellitus gestasional (GDM) memiliki dampak pada pola pertumbuhan keturunan. [8] Penelitian ini dilakukan oleh Sharma *et al.* (2019) bertujuan untuk menyelesaikan masalah pendeteksian dan prediksi diabetes menggunakan teknik machine learning dan internet of things (IOT). Metode yang digunakan studi survei yang mencakup tinjauan terhadap literatur dan penelitian terkait dalam memahami status terkini penelitian dalam menentukan diabetes dan kerangka kerja yang diusulkan.[9] Perbedaan fokus penelitian ini berbeda, tidak berfokus pada deteksi atau prediksi diabetes, melainkan pada dampaknya terhadap pola pertumbuhan keturunan. Penelitian yang saya berikan berfokus pada penggunaan algoritma naïve bayes dalam pendekatan data mining untuk deteksi diabetes mellitus, sedangkan pada penelitian ini juga deteksi penyakit diabetes mellitus digunakan untuk mendiagnosis secara mandiri sebelum pergi ke dokter. [10] Tujuan dari analisis data mining naïve bayes untuk deteksi diabetes mellitus adalah untuk meningkatkan pemahaman tentang faktor-faktor yang mempengaruhi risiko diabetes serta untuk mengembangkan alat yang dapat membantu dalam deteksi dini, intervensi, dan manajemen penyakit.[11][12] Dengan menggunakan pendekatan ini, diharapkan dapat tercapai peningkatan dalam upaya pencegahan dan pengelolaan diabetes mellitus, serta mengurangi beban yang ditimbulkan oleh penyakit ini bagi individu dan masyarakat secara keseluruhan

2. METODE PENELITIAN

Dalam studi ini, algoritma Naïve Bayes digunakan sebagai metode untuk mengukur tingkat akurasi dalam mengklasifikasikan penyakit diabetes mellitus[13].

A. Sumber Data

Penelitian ini memanfaatkan data yang diperoleh dari dataset yang tersedia di situs web Kaggle. Dataset yang digunakan adalah *Early Stage Diabetes Risk Prediction*, dengan nama file *diabetes.csv*. Penelitian ini melibatkan 9 variabel dan mencakup total 899 data.

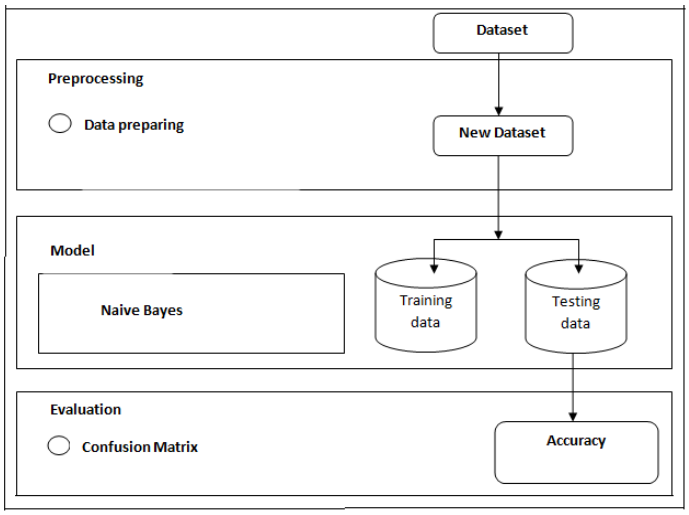
Dalam mendukung penelitian ini, dilakukan proses pengumpulan data. Data yang digunakan merupakan data sekunder yang diperoleh dari dataset yang tersedia di platform Kaggle dengan judul *"diabetes datasets."* Data sekunder merujuk pada data yang sudah diolah dan dikumpulkan oleh pihak lain, sehingga peneliti tidak perlu melakukan pengumpulan data secara langsung di lapangan.

Atribut data, yang juga dikenal sebagai fitur, merupakan karakteristik khusus dari sebuah entitas data. Dalam *Data Science* atau klasifikasi, atribut adalah variabel yang memberikan informasi tentang catatan data atau contoh tertentu. Setiap atribut memiliki tipe data tertentu dan merepresentasikan satu aspek dari entitas yang dijelaskan. Rincian atribut dalam kumpulan data yang digunakan dalam penelitian ini disajikan pada Tabel 1.

Table 1. Data Attributes

No	*Atteribut*	*Descripsion*
1	*Pregnancies*	Menyatakan Jumlah kehamilan
2	*Glucose*	Kadar Glukosa dalam darah
3	*BloodPressure*	Menyatakan ketebalan kulit
4	*SkinThickness*	Menyatakan ketebalan kulit
5	*Insulin*	Memperlihatkan tingkat *Insulin* dalam darah
6	*BMI*	Menyatakan indeks massa tubuh
7	*Diabetes*	Persentase diabetes
8	*Age*	Menyatakan umur
9	*Outcome*	Menyatakan hasil akhir 1 adalah Ya dan 0 adalah Tidak

B. Kerangka Pemikiran



Gambar 1 Kerangka Penelitian

Tahap preprocessing data dilakukan dengan mengatasi nilai yang hilang menggunakan teknik mengabaikan tupel dengan data yang tidak lengkap. Setelah proses ini, total data yang tersisa berjumlah 899 instans, terdiri dari 86 data positif dan 516 data negatif. Dua variabel kelas digunakan untuk menentukan apakah seorang pasien berisiko menderita diabetes (positif) atau tidak (negatif) [14].

Selanjutnya, *RapidMiner* digunakan untuk melatih dan menguji dataset baru menggunakan algoritma Naïve Bayes. Analisis hasil mencakup metrik seperti kesalahan klasifikasi (*classification error*), akurasi probabilitas maksimal untuk setiap kelas, nilai *recall*, dan presisi. Dataset kemudian dievaluasi menggunakan matriks kebingungan (*confusion matrix*) untuk mengukur tingkat akurasi model[15].

C. Algoritma Naïve Bayes

Naïve Bayes adalah algoritma pembelajaran mesin yang umum digunakan dalam tugas klasifikasi, terutama dalam analisis data teks dengan dimensi tinggi. Contohnya mencakup analisis sentimen, penyaringan spam, dan berbagai klasifikasi lainnya. Algoritma ini dikenal karena kesederhanaan dan efektivitasnya, serta kemampuannya dalam membangun model dengan cepat, menjadikannya salah satu algoritma prediksi yang efisien. Nama "naïve" merujuk pada asumsi bahwa kemunculan suatu fitur tidak bergantung pada kemunculan fitur lainnya, meskipun dalam praktiknya, fitur-fitur tersebut dapat saling berkaitan.[16] Algoritma ini didasarkan pada teorema Bayes, yang menghitung probabilitas kondisional berdasarkan informasi awal yang tersedia. Kelebihan algoritma Naïve Bayes meliputi kecepatan dalam membangun model, kemampuan prediksi, dan pendekatan baru untuk memahami data. Namun, algoritma ini hanya mendukung atribut dengan tipe data diskrit atau yang telah diubah menjadi diskrit, sehingga tidak kompatibel dengan atribut bernilai kontinu. Selain itu, algoritma ini mengasumsikan bahwa setiap atribut bersifat independen dan memberikan kontribusi yang setara terhadap prediksi atribut target[17].

Rumus *Naive Bayes* didasarkan pada Teorema Bayes:

$$P(C/X) = \frac{P(X/C) \cdot P(C)}{P(X)} \quad (1)$$

- *P(C/X)* = probabilitas kelas C diberikan fitur X
- *P(X/C)* = probabilitas mendapatkan fitur X diberikan kelas C
- *P(C)* = probabilitas dari kelas C (*prior probability*)
- *P(X)* = probabilitas dari fitur X (*evidence*)

1. ***Preprocessing Data***:
   - Bersihkan data (tangani *missing values*, hapus *noise*).
   - Jika data berupa teks, lakukan tokenisasi, *stemming*, dan konversi ke representasi numerik.

2. **Hitung *Probabilitas Prior*** *(P(C)):*
   - Probabilitas setiap kelas dihitung dari distribusi data pelatihan.

$$P(c) = \frac{\text{Jumlah data di kelas } C}{\text{Total data}} \quad (1)$$

3. **Hitung *Likelihood*** *(P(X/C)):*
   - Untuk fitur kategorikal, hitung frekuensi kemunculan fitur dalam kelas tertentu.
   - Untuk fitur kontinu, gunakan distribusi Gaussian:

$$P(X|C) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right) \quad (2)$$

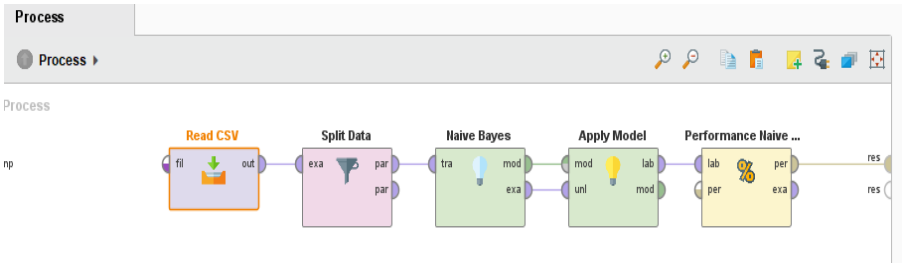
4. **Kalkulasi *Probabilitas Posterior*** *(P(C/X)):*
   - Gabungkan *prior* dan *likelihood*:

$$P(C|X) \propto P(C) \cdot \prod_{i=1}^{n} P(x_i|C) \quad (3)$$

5. **Prediksi**:
   - Pilih kelas *C* dengan probabilitas *P(C/X))* tertinggi. (4)

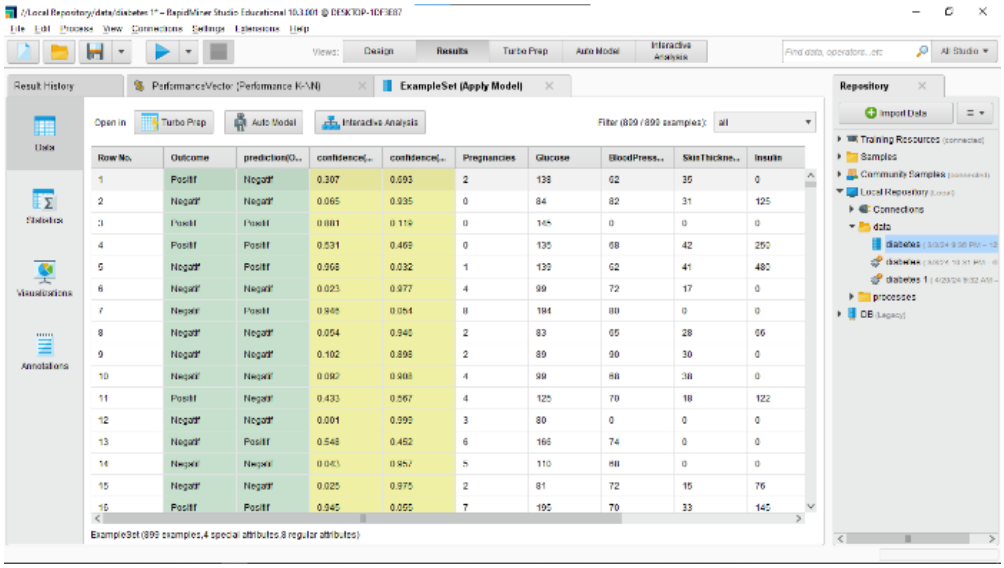
3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, dilakukan percobaan dengan mengimpor data dari situs resmi kaggle dalam format excel. Data kemudian diproses dengan membaginya menjadi dataset terpisah sebelum akhirnya menjalankan algoritma naïve bayes. Langkah berikutnya adalah menerapkan model untuk mengolah data dan akhirnya menghasilkan hasil, sebagaimana terlihat pada gambar 2 :



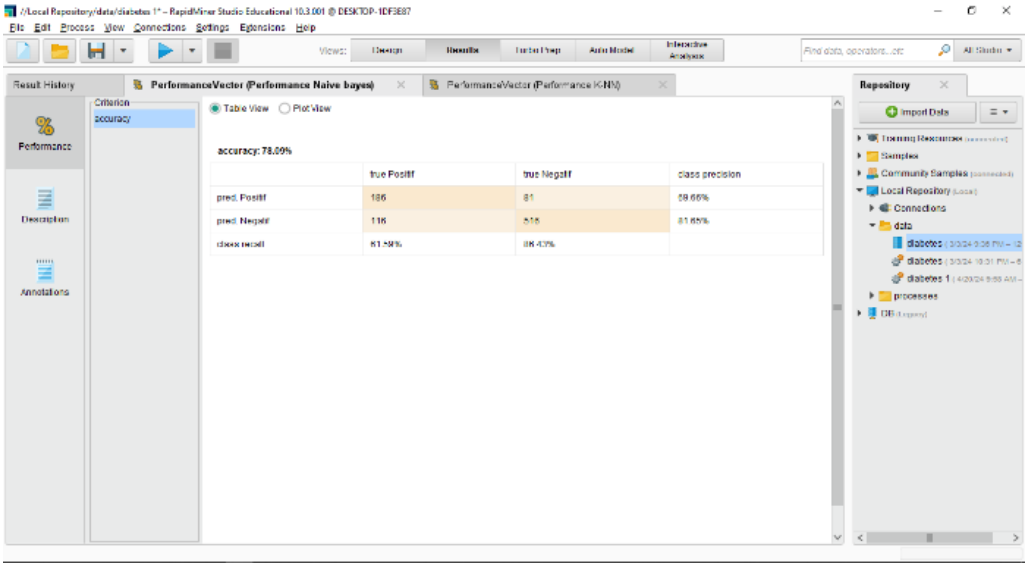
Gambar 2. Proses Aplode Dataset

Setelah data di masukan disini bisa kita lihat terdapat beberapa kolom yang terdiri dari Outcome, Prediction dan lain-lain seperti pada gambar 3:



Gambar 3. Proses Testing Dataset

Dari hasil percobaan di atas didapatkan skor tertinggi dengan jumlah data train 70% dan data testing 30%. Didapatkan akurasi sebesar 78.09 % seperti pada gambar 4:



Gambar 4. Akurasi Naïve Bayes

Accuracy tertinggi pada metode naïve bayes di dapat 78.09% dengan imputation missing value menggunakan mean. Hasil dari evaluasi pengklasifikasian dengan naïve bayes menghasilkan seperti gambar 5:

```
PerformanceVector:
accuracy: 78.09%
ConfusionMatrix:
True:   Positif Negatif
Positif:        186     81
Negatif:        116     516
```

Gambar 5. PerformanceVector Naïve Bayes.

4. KESIMPULAN

Dari hasil penelitian tersebut, terungkap bahwa metode naïve bayes menghasilkan tingkat akurasi tertinggi sebesar 78% dengan menerapkan metode imputasi missing value menggunakan mean. Pada tahap pembersihan data, teridentifikasi beberapa data yang tidak lengkap. Hasil menunjukkan bahwa dengan menggunakan metode imputasi missing value menggunakan median, metode klasifikasi Naïve Bayes mencapai tingkat akurasi sebesar 78.09%. Diharapkan penelitian ini dapat mendukung upaya pencegahan dan pengelolaan diabetes mellitus, serta membantu mengurangi beban bagi individu dan masyarakat secara keseluruhan.

DAFTAR PUSTAKA

[1] U. M. Butt *et al.*, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," vol. 2021, 2021.

[2] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big Data*, 2019, doi: 10.1186/s40537-019-0175-6.

[3] A. A. Utomo, S. Rahmah, and R. Amalia, "faktor risiko diabetes mellitus tipe 2 :," vol. 01, pp. 44–53, 2020.

[4] C. Wu, L. Huang, F. Chen, C. Kuo, and D. Yeih, "Using Machine Learning to Predict Abnormal Carotid Intima-Media Thickness in Type 2 Diabetes," pp. 1–13, 2023.

[5] L. F. Khaerunnisa and A. Fajarwati, "Pengawasan Penerimaan Peserta Didik Baru Jalur Keluarga Ekonomi Tidak Mampu," *J. Agreg. Aksi Reformasi Gov. dalam Demokr.*, vol. 7, no. 2, pp. 163–176, 2019, doi: 10.34010/agregasi.v7i2.2559.

[6] K. Puspita, Y. Alkhalifi, and H. Basri, "Rancang Bangun Sistem Informasi Penerimaan Peserta Didik Baru Berbasis Website Dengan Metode Spiral," *Paradig. - J. Komput. dan Inform.*, vol. 23, no. 1, pp. 35–42, 2021, doi: 10.31294/p.v23i1.10434.

[7] T. Madhubala, R. Umagandhi, and P. Sathiamurthi, "Diabetes Prediction using Improved Artificial Neural Network using Multilayer Perceptron," vol. 9, no. 12, pp. 167–179, 2022.

[8] S. Shafaeizadeh, L. Harvey, M. Abrahamse-berkeveld, and L. Muhardi, "Gestational Diabetes Mellitus Is Associated with Age-Specific Alterations in Markers of Adiposity in O ff spring : A Narrative Review," pp. 1–10.

[9] R. Asrianto and M. Herwinanda, "Jurnal Computer Science and Information Technology (CoSciTech) algoritma support vector machine," vol. 3, no. 3, pp. 431–440, 2022.

[10] M. Jannah, E. Erawan, and H. Burhanuddin, "Implementasi Program Penerimaan Peserta Didik Baru (Ppdb) Online Di Smp Negeri 21 Samarinda," *Ejournal.Ap.Fisip-Unmul.Ac.Id*, vol. 8, no. 3, pp. 9303–9317, 2020, [Online]. Available: https://ejournal.ap.fisip-unmul.ac.id/site/wp-content/uploads/2020/08/EJOURNAL B (08-05-20-05-29-34).pdf

[11] I. N. Sanita, S. Defit, and G. W. Nurcahyo, "Jurnal Computer Science and Information Technology (CoSciTech) Sistem Pendukung Keputusan Menggunakan Metode Multi Attribute Utility Theory (MAUT) Decision Support System Using The Method Attribute Utility Theory (MAUT) For Digital Service Selection," vol. 4, no. 1, pp. 216–225, 2023.

[12] M. Unik and Sri Nadriati, "Overview: Random Forest Algorithm for PM2.5 Estimation Based on Remote Sensing," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 3, pp. 422–430, 2022, doi: 10.37859/coscitech.v3i3.4380.

[13] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," vol. IV, no. September, pp. 15–21, 2020.

[14] A. Afifuddin and L. Hakim, "Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4 . 5," vol. 3, no. September, pp. 25–33, 2023.

[15] R. H. Tanjung *et al.*, "Jurnal Computer Science and Information Technology (CoSciTech)," vol. 4, no. 1, pp. 193–199, 2023.

[16] D. S. Simatupang and S. Nursinta, "Jurnal Computer Science and Information Technology (CoSciTech) Sentiment Analysis of Job Vacancy Hoax News Using The Naive Bayes Method," vol. 5, no. 2, pp. 474–482, 2024.

[17] M. A. Wiratama and W. M. Pradnya, "optimasi algoritma data mining menggunakan backward elimination untuk klasifikasi penyakit diabetes jurnal nasional pendidikan teknik informatika : janapati | 2," vol. 11, pp. 1–12, 2022.