P- ISSN: 2355-6684, E-ISSN: 2776-6640
https://ojs.unpkediri.ac.id/index.php/noe

IMPLEMENTASI METODE NAIVE BAYES UNTUK KLASIFIKASI RISIKO PENYAKIT HIPERTENSI

Muhammad Nafid Zanis[1*], Ghufron[2]
Universitas Islam Sultan Agung Semarang[1,2]
nafidzanis6@std.unissula.ac.id[1], ghufron@unissula.ac.id[2]
Corresponding author

Abstrak

Hipertensi merupakan gangguan kesehatan kronis yang ditunjukkan oleh peningkatan tekanan darah dalam arteri dan sering kali tidak disadari oleh penderitanya karena minimnya gejala awal, sehingga disebut sebagai silent killer. Faktor penyebabnya meliputi gaya hidup tidak sehat, seperti stres, obesitas, merokok, konsumsi alkohol, makanan berlemak tinggi, serta faktor usia dan keturunan. Seiring perkembangan teknologi, metode data mining digunakan untuk membantu proses deteksi penyakit, salah satunya melalui teknik klasifikasi. Penelitian ini memanfaatkan algoritma Naïve Bayes untuk mengklasifikasikan tingkat risiko hipertensi. Berdasarkan pengujian terhadap 3.000 data latih dan 751 data uji, diperoleh hasil yang cukup baik yaitu akurasi 87%, presisi 90%, recall 91%, dan F1-score 91%. Hasil yang diperoleh menunjukkan bahwa penerapan metode Naïve Bayes mampu melakukan klasifikasi dengan performa yang tinggi.

Kata Kunci : *Data Mining*, *Hipertensi, Klasifikasi, Naïve Bayes*

A. PENDAHULUAN

Tekanan darah tinggi atau hipertensi adalah gangguan kronis yang terjadi ketika tekanan darah dalam arteri meningkat secara terus-menerus. Berdasarkan panduan dari *American Heart Association* (AHA), seseorang diklasifikasikan sebagai pengidap penyakit hipertensi apabila tekanan darah sistoliknya tercatat di angka ≥130 mmHg atau tekanan darah diastoliknya tercatat di angka ≥80 mmHg. Hipertensi kerap disebut sebagai *silent killer* karena umumnya penderitanya tidak menyadari bahwa mereka mengidap kondisi tersebut sebelum dilakukan pemeriksaan darah. Kebiasaan hidup yang tidak sehat menjadi penyebab utama terjadinya hipertensi seperti stres, obesitas, merokok, konsumsi alkohol, serta makanan yang mempunyai tinggi lemak. Faktor lain seperti usia, memiliki riwayat hipertensi dalam keluarga, jenis kelamin juga bisa mempengaruhi terjadinya hipertensi (Setiandari L.O, 2022).

Perkembangan teknologi yang semakin pesat membuat komunitas bidang medis terbantu dengan adanya teknologi yang mampu mendeteksi penyakit secara tepat dan akurat. Alternatif yang digunakan yaitu dengan memanfaatkan beberapa informasi data dan membuat model untuk melakukan prediksi penyakit hipertensi dengan teknik klasifikasi dalam data mining. Sebagai salah satu teknik data mining, klasifikasi bertujuan untuk pengelompokan nilai variabel yang belum diketahui dengan merujuk pada variabel-variabel yang sudah diketahui sebelumnya (Riany dan Testiana, 2023).

Penelitian sebelumnya dilakukan untuk mengetahui kinerja Naïve Bayes dan Random Forest pada klasifikasi dua kelas yaitu kategori ya dan tidak mendapatkan tingkat akurasi sebesar 85,90% pada metode Naïve Bayes. Hal ini menunjukkan metode yang digunakan mampu melakukan klasifikasi dengan baik (Kharits, Hayati dan Bahtiar, 2023). Penelitian yang lain untuk mengetahui perbandingan akurasi pada metode SVM dengan parameter Grid Search dan GA memperoleh hasil akurasi sebesar 89,42%. Hasil tersebut menunjukkan bahwa metode yang diterapkan terbukti efektif dalam mengklasifikasikan penyakit hipertensi (Awalullaili, Ispriyanti dan Widiharih, 2022). Penelitian sebelumnya yang betujuan melakukan prediksi pada jenis gangguan penyakit hipertensi menggunakan metode ANN mendapatkan tingkat akurasi sebesar 85% (Purwono *dkk.*, 2022). Penelitian lain yang bertujuan untuk melakukan klasifikasi pada status pasien pengidap hipertensi dengan metode Regresi Logistik Biner dan Naïve Bayes Classifier mendapat akurasi tertinggi pada proporsi 90:10 sebesar 73,08% pada model naïve bayes (Sahila, Widiharih dan Utami, 2024). Penelitian sebelumnya yang dilakukan untuk mengetahui tingkat akurasi yang optimal dengan metode *Decision Tree* dan *Random Forest* mendapat akurasi sebesar 100% (Aditya, Lutvi Azizah dan Indahyanti, 2024).

Sebagaimana telah diuraikan dalam penelitian sebelumnya mengenai latar belakang masalah, penulis tertarik untuk melakukan klasifikasi tingkat risiko penyakit hipertensi pada pasien dengan tujuan mengetahui akurasi pada data menggunakan metode Naïve Bayes dengan perluasan dataset.

B. LANDASAN TEORI

1. Klasifikasi

   Klasifikasi merupakan sebuah teknik dalam pengolahan data yang bertujuan untuk mengelompokkan data berdasarkan kriteria yang sudah diterapkan. Klasifikasi memainkan peran penting dengan menentukan tingkat akurasi data yang diperoleh dengan memanfaatkan berbagai metode yang tersedia (Akmal, Faqih dan Dikananda, 2023).

2. Hipertensi

   Hipertensi merupakan gangguan penyakit kronis yang ditandai oleh meningkatnya tekanan darah di arteri yang dapat berkembang secara perlahan tanpa disadari. Kondisi ini diukur dengan mempertimbangkan dua parameter, yaitu sistolik yang muncul ketika jantung berkontraksi dan diastolik yang terjadi saat jantung berelaksasi antara dua denyut. Hipertensi sering kali dijuluki sebagai *sillent killer* karena pada tahap awal tidak menimbulkan gejala yang mencolok. Namun jika tidak segera ditangani, hipertensi dapat menyebabkan komplikasi yang cukup serius. Komplikasi yang umum terjadi mencakup penyakit kardiovaskular seperti stroke dan gagal jantung, kerusakan ginjal, gangguan penglihatan, serta aneurisma (Tarimana *dkk.*, 2024).

3. Data Mining

   Metode data mining digunakan untuk mengekstraksi informasi tertentu dari sebuah *database*. Data mining dapat diterapkan melalui berbagai teknik seperti, *artificial intelligence*, *machine learning,* dan metode statistika yang bermanfaat dalam mengetahui informasi bentuk data dalam skala besar (Yoliadi, 2023). Data mining merupakan tahapan penting dalam proses KDD (*Knowledge Discovery Database* yang berperan dalam mengekstraksi pola bermakna dari kumpulan data besar yang tersimpan dalam basis data (Zai, 2022). KDD terdiri atas beberapa langkah sistematis seperti pengumpulan data, praproses untuk membersihkan data mentah, transformasi data, dan terakhir evaluasi (Kartika *dkk.*, 2022).

4. Naïve Bayes

   Algoritma Naïve Bayes merupakan salah satu teknik dalam metode *machine learning* yang digunakan dalam proses klasifikasi data. Menurut (Siska *dkk.*, 2023) perhitungan metode Naïve Bayes menggunakan persamaan yang tercantum dalam rumus (1):

$$P\,(A|E) = \frac{P\,(E|A)\,.\,P(A)}{P(A)} \quad (1)$$

   Catatan dari rumus di atas sebagai berikut :

   $P(A|E)$: Nilai peluang terjadinya A apabila E diketahui

   $P(E|A)$: Nilai peluang terjadinya E apabla A diketahui

   $P(A)$: Nilai peluang kejadian A

   $P(E)$: Probabilitas kejadian E

   Selanjutnya, jika terdapat atribut yang memiliki nilai kontinu, perhitungan nilai posterior dilakukan dengan cara yang berbeda (Wie dan Siddik, 2022). Atribut yang memiliki nilai kontinu diasumsikan mengikuti distribusi Gaussian dengan menggunakan *mean* dan *standar deviasi* yang ditunjukkan pada persamaan (2) :

$$P\,(A) = \frac{1}{\sqrt{2\pi\sigma}}\,e - \frac{(A-\mu)^2}{2\sigma^2} \quad (2)$$

   Keterangan dari persamaan rumus di atas adalah sebagai berikut :

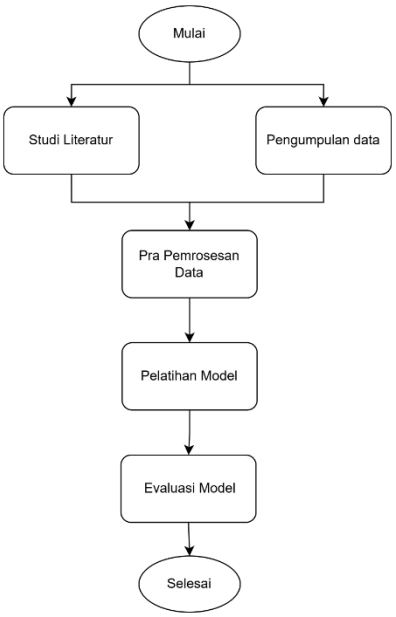
   $P(A)$ = peluang nilai A dalam distibusi normal

   e = bilangan euler

μ = nilai rata-rata setiap atribut
σ = nilai standar deviasi setiap atribut
π = konstanta pi

C. METODE PENELITIAN

Metode Naïve Bayes digunakan sebagai pendekatan klasifikasi untuk mengelompokkan risiko penyakit hipertensi. Penelitian ini memerlukan beberapa tahapan yang disajikan pada gambar :



Gambar 1. Flowchart Metode Penelitian

1. Studi Literatur
   Peneliti mencari berbagai sumber informasi mengenai teori yang mendasari serta berkaitan langsung dengan fokus penelitian ini. Teori yang dipelajari meliputi klasifikasi, data mining, Naïve Bayes, dan hipertensi yang bersumber dari artikel ilmiah, jurnal, buku, tugas akhir, serta situs *website* yang tersedia di situs internet.
2. Pengumpulan Data
   Data yang diambil merupakan data sekunder yang diperoleh dari situs *website* Kaggle.com dengan judul "*Hypertension Risk Model Main"*. Data tersebut diunggah oleh MD Raihan Khan pada tahun 2024.
3. Pra-Pemrosesan Data
   Pemrosesan data dilaksanakan dengan tujuan untuk memperbaiki kualitas data yang akan diteliti. Dalam penelitian ini, tahapan pemrosesan data yang akan dilakukan mencakup :
   a) Pembersihan data, yang bertujuan untuk mencegah terjadinya nilai yang hilang (missing value), dilakukan dengan cara menghapus atau menghilangkan seluruh data yang memiliki nilai yang hilang atau yang tergolong sebagai *null values* (Paramitha *dkk.*, 2023).
   b) Identifikasi terhadap nilai *outlier* yang dapat mempengaruhi distribusi data serta hasil analisis dengan cara mengganti nilai *outlier* agar tetap berada dalam rentang risiko rendah.
4. Pelatihan Model
   Untuk mengetahui kategori tingkat risiko penyakit hipertensi, data yang telah dilakukan pra-pemrosesan data pada tahap ini akan diimplementasikan dengan metode *Naïve Bayes*.
5. Evaluasi Model

Evaluasi terhadap metode Naïve Bayes menggunakan *confusion matrix* sebagai alat pengujian. *Confusion matrix* merupakan suatu tabel yang memberikan *output* kerja berdasarkan hasil prediksi.

D. HASIL DAN PEMBAHASAN

Kumpulan data yang diperoleh berjumlah 4.240 data yang terdiri dari 11 atribut dan 1 kelas meliputi variabel *sex* (jenis kelamin), *age* (umur), *currentSmoker* (riwayat merokok), *CigsPerDay* (jumlah batang rokok yang dihisap tiap hari), *BPMeds* (penggunaan obat darah tinggi), *totChol* (total kolesterol), *sysBP* (tekanan darah sistolik), *diaBP* (tekanan darah diastolik), *BMI* (indeks massa tubuh), *heartRate* (detak jantung), *Glucose* (kadar glukosa), dan label *Risk* (kelas risiko rendah atau hipertensi).

Tabel 1. Rangkuman dari Atribut Dataset

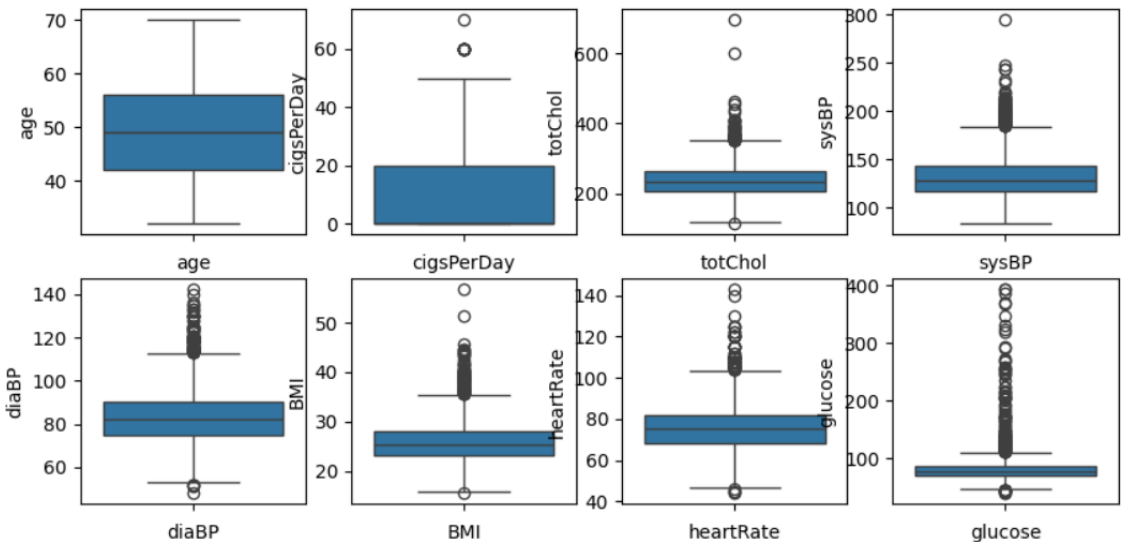
No	Atribut	Nilai
1	*Sex*	0, 1
2	*Age*	32-70
3	*currentSmoker*	0, 1
4	*cigsPerDay*	0-20
5	*BPMeds*	0, 1
6	*Diabetes*	0, 1
7	*totChol*	107-696
8	*SysBP*	83.5-295
9	*DiaBP*	48-143
10	*BMI*	15.54-56.80
11	*heartRate*	44-143
12	*Glucose*	40-394
13	*Risk*	0, 1

1. Pra-Pemrosesan Data

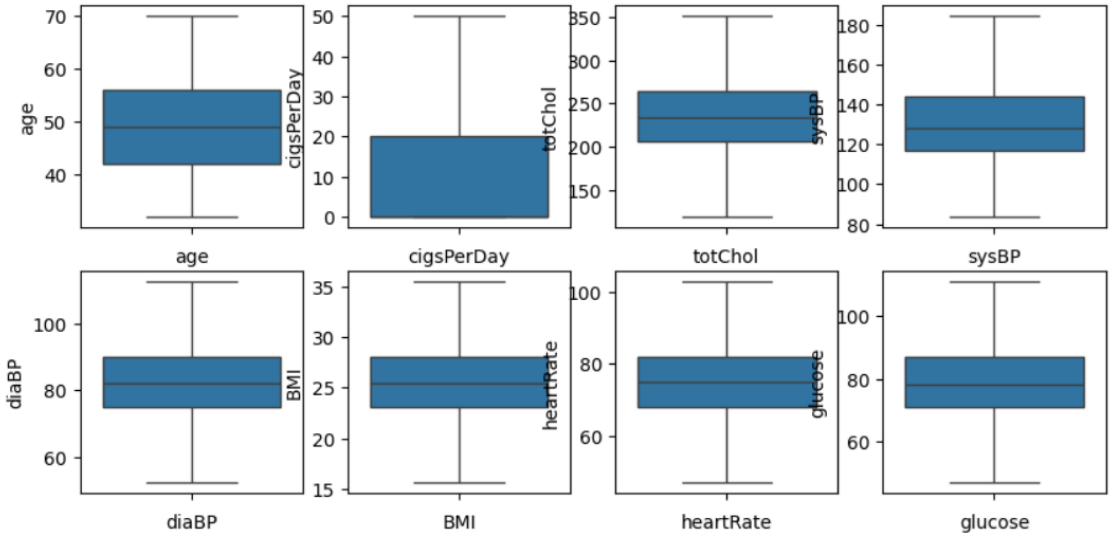
a) Pembersihan Data

Pada tahap ini, dataset rekam medis akan dibersihkan untuk menghilangkan data yang memiliki nilai hilang. Terdapat sejumlah 29 data *missing value* pada atribut cigsPerDay, 53 data pada atribut BPMeds, 50 data pada atribut totChol, 19 data pada atribut BMI, 1 data pada atribut heartRate, dan 388 data pada atribut glucose. Setelah proses pembersihan data, jumlah data yang tersisa menjadi 3.751 data.

b) Mengecek nilai *outlier*



Gambar 2. Mengecek Nilai *Outlier*



Gambar 3. Atribut setelah di *Outlier*

2. Pelatihan Model

a) Menentukan data pelatihan dan data pengujian

Tabel 2. Membagi Dataset *Training* dan *Testing*

Keterangan	Rasio (%)	Jumlah Data
Data Training	80%	3.000
Data Testing	20%	751
Total	**100%**	**3.751**

Pada tabel diatas menunjukkan keterangan yang dilakukan dalam pelatihan model dengan membagi dataset menjadi dua bagian, yaitu data latih dan data uji yang diatur dalam rasio perbandingan data latih yang digunakan sebanyak 80% dengan total 3.000 data sedangkan data uji sebanyak 20% dengan total 751 data.

b) Menghitung probabilitas atribut

Tahap pertama dalam perhitungan metode Naïve Bayes adalah menghitung probabilitas pada pasien yang termasuk kategori kelas risiko rendah atau risiko tinggi sebelum diimplementasikan ke dalam aplikasi.

$$P\ (jumlah\ kelas\ risiko\ rendah) = \frac{P\ (jumlah\ kelas\ risiko\ rendah)}{P\ (jumlah\ semua\ data)} = \frac{2065}{3.000} = 0{,}6886$$

$$P\ (jumlah\ kelas\ risiko\ tinggi) = \frac{P\ (jumlah\ kelas\ risiko\ tinggi)}{P\ (jumlah\ semua\ data)} = \frac{935}{3.000} = 0{,}3114$$

Kemudian, menghitung nilai probabilitas pada atribut yang memiliki tipe kategorikal seperti atribut *sex, current Smoker, BPMeds*, dan diabetes.

Tabel 3. Menghitung Probabilitas Atribut

Atribut	Sub Atribut	Probabilitas	
		Risiko rendah	Risiko tinggi
Sex	*Male*	0,55061	0,54866
	Female	0,44939	0,45134
currentSmoker	*No*	0,47264	0,58182
	Yes	0,52736	0,41818
BPMeds	*No*	0,52397	0,51765
	Yes	0,47603	0,48235
diabetes	*No*	0,48959	0,50053

P- ISSN: 2355-6684, E-ISSN: 2776-6640
https://ojs.unpkediri.ac.id/index.php/noe

Yes	0,51041	0,49947

Selanjutnya, menghitung *mean* dan *standar deviasi* dari atribut yang memiliki tipe numerik.

Tabel 4. Menghitung *mean*

Mean								
Label	***age***	***cigsPerDay***	***totChol***	***sysBP***	***diaBP***	***BMI***	***heartRate***	***glucose***
0	47,685	9,703	231,580	121,762	78,137	7,0372E+13	74,185	79,0712
1	53,585	7,866	247,976	150,591	92,026	242,071	77,987	81,551

Tabel 5. Menghitung *standar deviasi*

Standar Deviasi								
Label	***age***	***cigsPerDay***	***totChol***	***sysBP***	***diaBP***	***BMI***	***heartRate***	***glucose***
0	8,1703	11,8660	41,1177	12,3336	8,0787	1,0485,E+15	10,9152	12,3935
1	8,0733	11,7122	42,9199	15,0431	9,2108	81,4661	11,8887	13,7840

Pengujian dilakukan dengan menggunakan sampel dari salah satu pasien dengan menerapkan perhitungan Naïve Bayes Gaussian.

	male	age	rrentSmok	cigsPerDay	BPMeds	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	Risk
Data uji ke	1	43	1	35	0	0	207	117	65	24,42	60	100	2,72E-17
0	0,449395	0,118439	0,527361	0,011939	0,523971	0,489588	0,052048	0,105465	0,037427	1,23E-08	0,051916	0,027241	3,07E-19
1	0,451337	0,059459	0,418182	0,007966	0,517647	0,500535	0,038616	0,008503	0,001776	0,001246	0,036846	0,043886	2,72E-17

Gambar 4. Pengujian dengan data latih

Pada gambar dilakukan perhitungan Gaussian Naïve Bayes pada sampel salah satu pasien. Dari perhitungan menunjukkan bahwa nilai probabilitas resiko rendah 3,07E-19 dan probabilitas risiko tinggi 2,72E-17 yang artinya nilai probabilitas pada risiko tinggi lebih besar daripada nilai probabilitas kategori risiko rendah. Oleh karena itu, berdasarkan hasil diatas, pasien yang memiliki nilai atribut tersebut diprediksi memiliki penyakit hipertensi dengan **Risiko Tinggi.**

3. Evaluasi Model

Tabel 5. *Confusion Matrix*

		Prediksi	
		Risiko rendah	**Risiko tinggi**
Aktual	Risiko rendah	465	51
	Risiko tinggi	46	189

Pada gambar merupakan tabel *confusion matrix* dari pengujian yang menggunakan metode Naïve Bayes. Dalam penelitian ini menggunakan percobaan dengan rasio data training 80% yang menghasilkan sebanyak 3.000 data latih dan 20% data training yang menghasilkan sebanyak 751 data latih, sehingga diperoleh rumus evaluasi model dari *confusion matrix* sebagai berikut :

Akurasi $= \frac{TP+TN}{TP+FP+FN+TN} \times 100\%$

$= \frac{465+189}{465+189+46+51} \times 100\% = 87\%$

Presisi $= \frac{TP}{TP+FP}$

$= \frac{465}{46+465} \times 100\% = 90\%$

Recall $= \frac{TP}{TP+FN}$

$$= \frac{465}{\quad} \text{ x } 100\% = 91\%$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{\quad}$$

$$= \frac{2 \times 90\% \times 91\%}{\quad} \text{ x } 100\% = 91\%$$

E. Kesimpulan dan Saran

Berdasarkan hasil penelitian yang telah dilakukan pada proses klasifikasi tingkat risiko penyakit hipertensi menggunakan metode Naïve Bayes, dapat diambil kesimpulan bahwa implementasi metode Naïve Bayes cukup akurat dalam melakukan klasifikasi dengan menggunakan pengujian 80% data training dengan *record* sebanyak 3.000 entri data dan 20% dilakukan untuk pengujian dengan total 751 data memberi hasil akurasi sebesar 87%, presisi sebesar 90%, *recall* sebesar 91% dan *F1-Score* sebesar 91%. Saran dari penelitian selanjutnya adalah menggunakan metode yang lain untuk perbandingan akurasi model pada data atau menerapkan SMOTE yang bertujuan menangani label *imbalanced data* agar menghasilkan akurasi yang lebih baik.

DAFTAR PUSTAKA

Aditya, M. F. R., Lutvi Azizah, N. dan Indahyanti, U. (2024) "Prediksi Penyakit Hipertensi Menggunakan Metode Decison Tree dan Random Forest," *Jurnal Ilmiah Komputasi*, 23(1), hal. 9–16.

Akmal, K., Faqih, A. dan Dikananda, F. (2023) "Perbandingan Metode Algoritma Naïve Bayes Dan K-Nearest Neighbors Untuk Klasifikasi Penyakit Stroke," *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(1), hal. 470–477.

Awalullaili, F. O., Ispriyanti, D. dan Widiharih, T. (2022) "Klasifikasi Penyakit Hipertensi Menggunakan Metode Svm Grid Search dan Svm Genetic Algorithm(GA)," *JURNAL GAUSSIAN*, 11(4), hal. 488–498.

Kartika, Y., Komariah, K., Surip, A., Saputra, R., & Ali, I. (2022). Implementasi Algoritma Naïve Bayes Untuk Prediksi Persediaan Barang Rotan. *Jurnal Ilmiah Informatika dan Komputer*, *4*(1), 33-40.

Kharits, A. K., Hayati, U. dan Bahtiar, A. (2023) "Perbandingan Prediksi Penyakit Hipertensi Menggunakan Metode Random Forest Dan Naïve Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(1), hal. 498–504.

Paramitha, N. Y., Nuryaman, A., Faisol, A., Setiawan, E., & Nurvazly, D. E. (2023). Klasifikasi Penyakit Stroke Menggunakan Metode Naïve Bayes. *Jurnal Siger Matematika*, 11-16.

Purwono, P., Dewi, P., Wibisono, S. K., & Dewa, B. P. (2022). Model Prediksi Otomatis Jenis Penyakit Hipertensi dengan Pemanfaatan Algoritma Machine Learning Artificial Neural Network. *Insect (Informatics and Security): Jurnal Teknik Informatika*, *7*(2), 82-90.

Riany, A. F. dan Testiana, G. (2023) "Penerapan Data Mining Untuk Klasifikasi Penyakit Jantung Koroner Menggunakan Algoritma Naive Bayes," *PROCEEDING Multi Data Palembang Student Conference*, 2(1), hal. 297–305.

Sahila, R., Widiharih, T. dan Utami, I. T. (2024) "Analisis Klasifikasi Menggunakan Regresi logistik Biner Dan Algoritma Naïve Bayes Classifier pada Penyakit Hipertensi," *JURNAL GAUSSIAN*, 13(2007), hal. 319–327.

Setiandari L.O, E. (2022) "Hubungan Pengetahuan, Pekerjaan dan Genetik (riwayat hipertensi dalam keluarga) Terhadap Perilaku Pencegahan Penyakit Hipertensi," *Media Publikasi Promosi Kesehatan Indonesia*, 5(4), hal. 457–462.

Siska, S. *dkk.* (2023) "Implementasi Metode Naive Bayes pada Prediksi Penyakit Seliak," *KOPERTIP*

Jurnal Ilmiah Manajemen Informatika dan Komputer, 7(1), hal. 8–13.

Tarimana, A. A. Z., Fajar, M. R. S., Saktiawan, M. A., & Saputra, R. A. (2024). PREDIKSI PENYAKIT HIPERTENSI MENGGUNAKAN MACHINE LEARNING DENGAN ALGORITMA REGRESI LOGISTIK. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(6), 12062-12068.

Wie, J. V. dan Siddik, M. (2022) "Penerapan Metode Naïve Bayes Dalam Mengklasifikasi Tingkat Obesitas Pada Pria," *JOISIE Journal Of Information System And Informatics Engineering*, 6(2), hal. 69–77.

Yoliadi, D. N. (2023) "Data Mining Dalam Analisis Tingkat Penjualan Barang Elektronik Menggunakan Algoritma K-Means," *Insearch: Information System Research Journal*, 3(01).

Zai, C. (2022) "Implementasi Data Mining Sebagai Pengolahan Data," *Jurnal Portal Data*, 2(3), hal. 1–12.