Institut Riset dan Publikasi Indonesia (IRPI)
SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat
Homepage: https://journal.irpi.or.id/index.php/sentimas
Prosiding Seminar Nasional/ ISSN(E): 2830-3083
03 Agustus 2023/ Pages: 49-55

Comparison of Naive Bayes and C4.5 Classification for Stoke Disease Diagnosis

Perbandingan Klasifikasi Naive Bayes dan C4.5 untuk Diagnosa Penyakit Stroke

Nadila Gusrialni Fitri[1*], Shofika Adilya[2], Apriliani[3] , Febryan Azizi[4]

[1,2,3,4] Program Studi Sistem Informasi, Fakultas Sains dan Teknologi,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia

E-Mail: [1]12050322581@students.uin-suska.ac.id [2]12050313116@students.uin-suska.ac.id,
[3]12050320412@students.uin-suska.ac.id [4]12050312098@students.uin-suska.ac.id

Corresponding Author: Nadila Gusrialni Fitri

Abstract

The disorder known as a stroke is caused by narrowed blood vessels in the brain, which can restrict or stop the flow of blood and oxygen to the brain. The purpose of this study is to compare the accuracy values of the C4.5 and Naive Bayes algorithms. The method taken in this study began with an understanding of the literature to gather theoretical foundations, as well as learning about stroke or cerebrovascular disease. Data is then collected which is obtained from the Kaggle website. The collected data is pre-processed, which consists of cleaning and transforming the data. Next, a model was created to evaluate and validate the accuracy of the C4.5 and Naïve Bayes algorithms on RapidMiner tools. The results showed that the C4.5 algorithm has an accuracy rate of 92.22% and the Naïve Bayes algorithm has an accuracy rate of 89.22%. It can be concluded that the C4.5 algorithm has a higher level of accuracy than the Naivee Bayes algorithm, so the C4.5 algorithm is the most optimal algorithm in the classification of stroke diagnosis.

Keyword: C4.5, Classification, Naïve Bayes, Stroke

Abstrak

Gangguan yang dikenal sebagai stroke ini disebabkan oleh penyempitan pembuluh darah di otak, yang dapat membatasi atau menghentikan aliran darah dan oksigen ke otak. Tujuan dari penelitian ini adalah untuk membandingkan nilai akurasi algoritma C4.5 dan Naive Bayes. Metode yang diambil dalam penelitian ini dimulai dengan tinjauan literatur untuk mengumpulkan dasar-dasar teori yang, serta mempelajari tentang penyakit stroke atau cerebrovascular. Data kemudian dikumpulkan yang diperoleh dari situs Kaggle. Data yang dikumpulkan diproses sebelumnya, yang terdiri dari pembersihan dan transformasi data. Selanjutnya, model dibuat untuk mengevaluasi dan memvalidasi akurasi algoritma C4.5 dan Naïve Bayes pada tools RapidMiner. Hasil penelitian menunjukkan bahwa algoritma C4.5 memiliki tingkat accuracy sebesar 92.22% dan algoritma naïve bayes dengan tingkat akurasi sebesar 89.22%. Dapat disimpulkan bahwa algoritma C4.5 memiliki tingkat akurasi yang lebih besar dibandingkan algoritma naivee bayes, maka algoritma C4.5 merupakan algoritma yang paling optimal dalam perbandingan klasifikasi diagnosa penyakit stroke.

Kata Kunci: C4.5, Klasifikasi, Naïve Bayes, Stroke

1. PENDAHULUAN

Gangguan yang dikenal sebagai stroke ini disebabkan oleh penyempitan pembuluh darah di otak, yang dapat membatasi atau menghentikan aliran darah dan oksigen ke otak. Sistem saraf dapat tersumbat, suplai darah dan oksigen dapat rusak atau terputus, dan organ yang terhubung ke sistem saraf dapat mengalami gangguan. [5].

Sumber penyakit stroke yang diduga dapat meningkatkan jumlah pengidap yaitu faktor makanan, stress serta gaya hidup, yang ditemukan pada pengecekan lemak darah pengidap. dalam riset yang dikerjakannya pada salah satu rumah sakit di Yogyakarta berpendapat bahwa kenaikan ataupun penurunan kolesterol bukan faktor efek pemicu terkena stroke [12].

Salah satu dari 10 penyakit dengan angka kematian tinggi di Indonesia adalah stroke. Hal ini berdasarkan data yang dikumpulkan dari sampel kematian yang representatif di Indonesia yang terjadi pada tahun 2014, berjumlah 41.590. Untuk setiap kematian ini, otopsi dilakukan sesuai dengan pedoman Organisasi Kesehatan Dunia secara real time oleh profesional medis atau staf terlatih. [10].

Pada pengidap pasca stroke umumnya ditemukan indikasi sisa akibat guna otak yang tidak membaik seluruhnya. Sebagian diantaranya yaitu kelumpuhan pada satu sisi badan, hilang rasa ataupun menyusutnya, gangguan penyeimbang, gangguan koordinasi, kendala berbahasa bahkan gangguan mental. Gangguan Fisik pada pengidap pasca stroke yaitu hemiparise(kelemahan satu sisi tubuh), ataupun hemiplegia(kelumpuhan pada satu sisi badan) dari satu bagian badan seperti wajah, lengan serta tungkai. Hal tersebut dapat menyebabkan penyusutan rentang gerak, kendala berbicara serta penyusutan kegiatan sehari-hari [9].

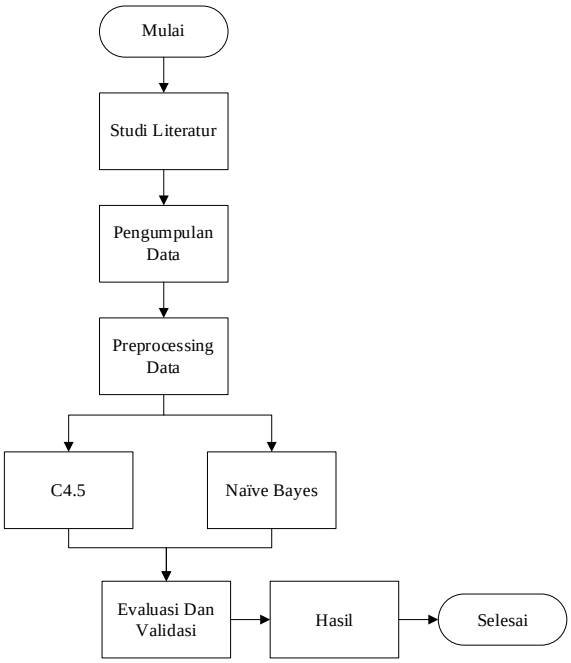
Proses menghubungkan setiap pola dalam setiap kumpulan data yang sangat besar dengan volume data yang sangat besar dikenal sebagai data mining [7]. Menemukan pola dan kondisi dalam database yang sangat besar sehingga dapat digunakan untuk menginformasikan keputusan adalah tujuan dari data mining. Klasifikasi adalah teknik yang sering digunakan dalam data mining karena berusaha mengkategorikan suatu objek ke dalam kelas tertentu berdasarkan pola kelas tertentu [2]. Untuk menentukan proses klasifikasi dan Algoritma Naive Bayes dan C4.5 pada dataset stroke, peneliti melakukan perbandingan terhadap klasifikasi diagnosa penyakit stroke dengan menggunakan Algoritma Naive Bayes dan C4.5 sebagai pengambilan keputusan. perbandingan ini dilakukan untuk mengetahui tingkat keakuratan dari kedua algoritma tersebut yang lebih baik.

Berdasarkan pada penelitian sebelumnya oleh (Handayani et al., 2021) Dengan akurasi 93,26%, Naive Bayes adalah algoritma paling akurat dalam situasi ini, sedangkan C4.5 memiliki akurasi 93,22%. Naive Bayes memiliki AUC sebesar 0,833 dan termasuk dalam predikat klasifikasi baik, sedangkan C4.5 memiliki AUC sebesar 0,758 dan termasuk dalam predikat klasifikasi memadai. Uji t berpasangan menghasilkan hasil 0,841, yang menunjukkan bahwa tidak ada perbedaan mencolok antara klasifikasi kedua algoritme untuk status kelayakan donor darah [8].

Penelitian selanjutnya dilakukan oleh (Ardiansyah et al., 2021) menunjukkan bahwa metode C4.5 (situasi 4) mengungguli algoritma Naive Bayes (skenario 2) dalam kategorisasi diabetes, dengan tingkat akurasi 99,03%, presisi 100%, dan daya ingat 98,18% [1]. Referensi ini berfungsi sebagai dasar untuk penerapan algoritme klasifikasi stroke C4.5 dan Naive Bayes dalam penelitian ini. Algoritma dengan kinerja terbaik akan dipilih berdasarkan perbandingan ini.

2. BAHAN DAN METODE

Metodologi penelitian merupakan penjabaran alur proses yang dilakukan dalam penelitian ini. Metodologi penelitian dapat divisualisasikan pada Gambar 1.



Gambar 1. Metodologi Penelitian

Gambar 1 merupakan alur proses yang dilakukan di penelitian ini. Dimulai dengan studi literatur untuk mengumpulkan dasar teori yang dibutuhkan, dan mendapatkan pengetahuan tentang penyakit pembuluh darah otak atau stroke. Kemudian informasi dikumpulkan melalui website Kaggle. Langkah selanjutnya adalah preprocessing, yang melibatkan pembersihan dan modifikasi data yang diperoleh. Keakuratan algoritma C4.5 dan Naive Bayes dalam program RapidMiner kemudian dievaluasi dan divalidasi menggunakan model yang dikembangkan setelahnya.

2.1 Algoritma C4.5

Suatu pendekatan yang dapat digunakan untuk mengatasi masalah dengan kategori atribut dalam klasifikasi data disebut algoritma C4.5 [3]. Algoritma C4.5 diterapkan untuk membentuk pohon keputusan yang memuat aturan-aturan dalam klasifikasi ([7]. Algoritma C4.5 ini digunakan untuk membuat pohon keputusan [2]. Tujuan membangun sebuah pohon keputusan harus memilih Attribute sebagai akar, buat cabang untuk setiap nilai, pisahkan kasus di setiap cabang, ulangi proses hingga semua kasus di cabang memiliki kelas yang sama [4]. Berikut langkah-langkah pembuatan pohon keputusan menggunakan algoritma C4.5 [6] :

1) Sebagai root, pilih atribut. Menemukan atribut dengan nilai tertinggi di antara atribut yang sudah ada menjadi dasar pemilihannya sebagai root. Persamaan berikut digunakan untuk menentukan nilai perolehan maksimum:

$$Gain(S, A) = Entropy(S) = \sum_{i=0}^{n} \frac{|S_i|}{|S|} * Entropy\ (s_i) \quad (1)$$

Keterangan :
- S : himpunan kasus
- A : atribut
- N : jumlah partisi atribut A
- |Si| : jumlah kasus pada partisi ke-i
- |S| : jumlah kasus dalam S

Rumus berikut dapat digunakan untuk menentukan nilai entropi:

$$Entropy(S)\ -\sum_{i=1}^{n} -pi * pi \quad (2)$$

Keterangan :
- S : himpunan partisi
- n : jumlah partisi S
- pi : proporsi dari Si tehadap S

2) Buat cabang untuk setiap nilai.
3) Pisahkan kasus di setiap cabang.
4) Ulangi proses hingga semua kasus di cabang memiliki kelas yang sama

2.2 Naïve Bayes

Klasifikasi statistik yang dapat digunakan untuk memperkirakan kemungkinan pengungkapan kelas adalah algoritma Nave Bayes. Pendekatan kategorisasi data menggunakan teknik ini [13]. fase penerapan algoritma untuk meneliti dataset dalam sistem implementasi Naive Bayes. Persamaan Teorema Bayes diberikan dalam rumus di bawah ini [11].

$$P(x|y)\ = -\frac{P(x)=P(x)}{P(y)} \quad (3)$$

Keterangan:
- y : data dengan kelas yang tidak terklasifikasi
- x : hipotesis data y merupakan suatu kelas spesifik
- P(x|y) : kemungkinan berdasarkan kondisi x dan y (posteriori probability)
- P(x) : kemungkinan premis x (prior probability)
- P(y|x) : probabilitas y mengingat keadaan dalam hipotesis x
- P(y) : kemungkinan y

3. HASIL DAN PEMBAHSAN

3.1 Preprocessing Data

Penelitian yang dilakukan dengan teknik pre-proccesing dilakukan untuk mendapatkan data yang berkualitas tinggi. Validasi data, yang melibatkan penghilangan outlier, noise, titik data kosong, dan data yang tidak konsisten, dan data diskritisasi, yang melibatkan pemilihan fitur gradasi, keduanya termasuk dalam pendekatan [3]. Penelitian ini menggunakan dataset stroke dari Kaggle yang memiliki 5110 record data dengan 12 atribut. Data kemudian disiapkan untuk digunakan melalui pra-pemrosesan, yang meliputi tahap pembersihan dan transformasi. Terdapat 1767 record yang tidak dapat digunakan dan telah dimusnahkan sebagai bagian dari pembersihan data. 3.343 catatan terdiri dari data yang tersedia untuk penggunaan, seperti yang ditunjukkan pada Tabel 1.

Tabel 1. Dataset Stroke

No	id	gender	age	hypertension	heart disease	ever married	work type	…	stroke
1	9046	1	67	0	1	Yes	Private	…	Ya
2	51676	2	61	0	0	Yes	Self-employed	…	Ya
3	31112	1	80	0	1	Yes	Private	…	Ya
…	…	…	…	…	…	…	…	…	
5110	44679	2	44	0	0	Yes	Govt_job	…	Tidak

Dari transformasi data proses pengorganisasian data ke dalam kategori tertentu yang memiliki data lebih kompleks tanpa menghilangkan isi, sehingga sederhana untuk diproses. Berikut adalah contoh transformasi data yang didokumentasikan pada Tabel 2.

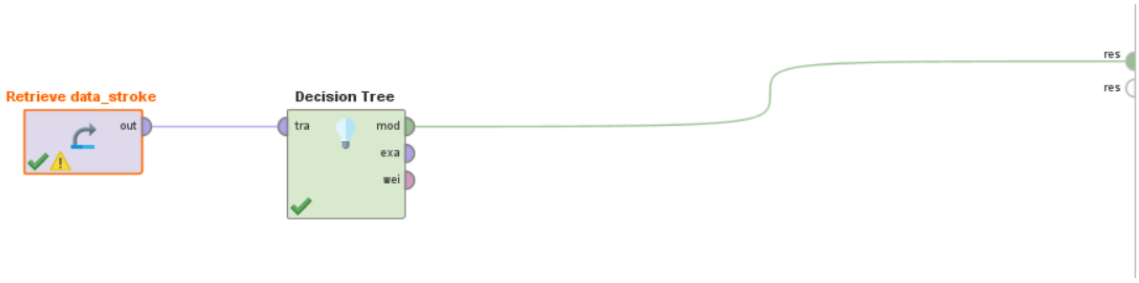
Tabel 2. Transformasi Data

No	id	gender	age	hypertension	heart disease	ever married	work type	…	stroke
1	9046	1	67	0	1	1	1	…	Ya
2	31112	1	80	0	1	1	2	…	Ya
3	60182	2	49	0	0	1	3	…	Ya
…	…	…	…	…	…	…	…	…	…
3343	37544	1	51	0	0	1	1	…	Tidak

3.2 Pembahasan

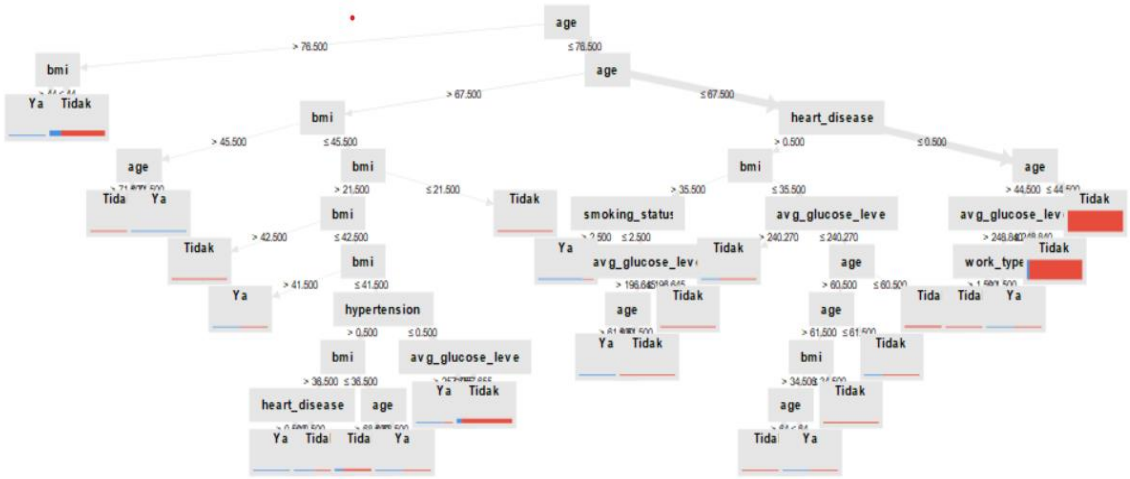
3.2.1 Implementasi Algoritma C4.5 dengan Tools Rapid Miner

Pada titik ini, dua operator digunakan yaitu Retrieve Data dan C4.5 yang memproses data menjadi informasi. Ambil Data digunakan untuk mengimpor data yang telah diimpor dari Excel.



Gambar 2. Klasifikasi Algoritma C4.5

Pohon keputusan yang dihasilkan dari model desain dataset diagnosis stroke menggunakan algoritma C4.5 ditunjukkan pada gambar di bawah.



Gambar 3. Decision Tree C4.5

Age adalah node tertinggi pada pohon keputusan pada gambar di atas, dengan kategori > 76.500, 6.500, dan 67.500.

3.2.2 Implementasi algoritma Naïve Bayes dengan Tools RapidMiner

Menggunakan Alat RapidMiner, metode Naive Bayes Menggunakan RapidMiner Tools, buat model algoritma Naive Bayes.

Gambar 4. Klasifikasi Naïve Bayes

Algoritma Nave Bayes digunakan pada gambar di atas sebagai model desain untuk pengumpulan data diagnosis stroke, dan aturan modelnya adalah sebagai berikut:

SimpleDistribution

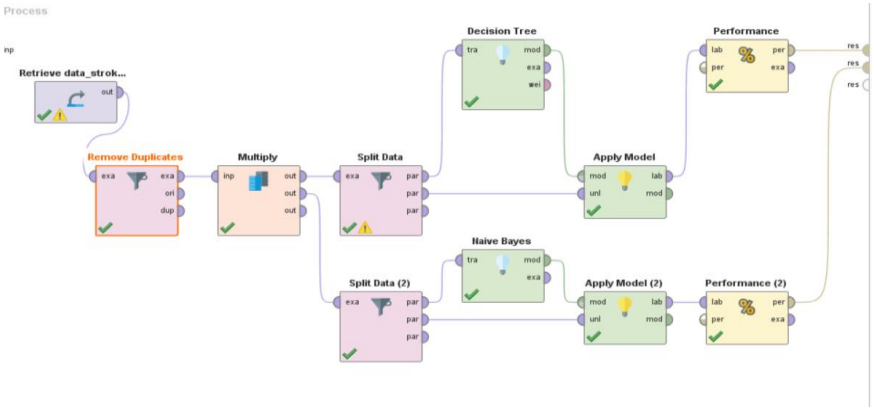
```
Distribution model for label attribute stroke


Class Ya (0.054)
9 distributions

Class Tidak (0.946)
9 distributions
```

Gambar 5. Hasil Algoritma Naïve Bayes

Hasil algoritma Naive Bayes ditunjukkan pada gambar di atas untuk kelas Ya dan Tidak masing-masing dengan nilai 0,054 dan 0,9464. Sehingga dapat dikatakan bahwa kelas No merupakan kelas tertinggi untuk diagnosis stroke.

3.3 Evaluasi dan Validasi

Menerapkan pemodelan Algoritma C4.5 berarti Naïve Bayes dapat dilihat pada Gambar 6.



Gambar 6. Menggambarkan Model Klasifikasi Algoritma C4.5 dan Nave Bayes

Contoh pelatihan yang digunakan sebagai operator algoritma C4.5, Naive Bayes, dan validasi pengujian ditunjukkan pada Gambar 6 sebagai tahap pertama dalam pengembangan model proses penyelesaian. Gambar ini juga menunjukkan cara mengambil dataset menggunakan RapidMiner. Karena akurasi 2 algoritma adalah sebagai berikut:

Tabel 3. Akurasi C4.5

Accurancy : 92.22%	true Ya	true Tidak	class precision
pred. Ya	13	55	19.12%
pred.Tidak	82	1610	95.15%
class recall	13.68%	96.70%	

Tabel 4. Akurasi Naïve Bayes

Accuracy : 89.22%	true Ya	true Tidak	class precision
pred. Ya	6	24	20.00%
pred.Tidak	12	292	96.05%
class recall	33.33%	92.41%	

3.4 Perbandingan hasil akurasi algoritma C4.5 dan Naïve Bayes

Hasil implementasi dengan tingkat akurasi antara Algoritma C4.5 dan teknik Naive Bayes.

Tabel 5. Perbandingan akurasi algoritma C4.5 dan Naive Bayes

No	Algoritma	Akurasi
1	C4.5	92.22%
2	Naïve Bayes	89.22%

4. KESIMPULAN

Sesuai hasil penelitian perbandingan algoritma C4.5 dan Naive Bayes pada pendataan diagnosa penyakit stroke menggunakan tools Rapidminer dengan algoritma C4.5 dengan accuracy 92.22% dan penggunaan prosedur pemecahan Naive Bayes dengan akurasi 89.22%. dapat disimpulkan bahwa algoritma C4.5 mempunyai tingkat accuracy yang lebih tinggi dibandingkan algoritma naive bayes, sebagai akibatnya algoritma C4.5 artinya prosedur pemecahan yang paling optimal dibandingkan menggunakan pembagian terstruktur mengenai diagnosis stroke.

REFERENSI

[1] Angraini, Y., Fauziah, S., & Putra, J. L. (2020). Analisis Kinerja Algoritma C4.5 Dan Naïve Bayes Dalam Memprediksi Keberhasilan Sekolah Menghadapi Un. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *5*(2), 285–290. https://doi.org/10.33480/jitk.v5i2.1233

[2] Ardiansyah, M., Sunyoto, A., & Luthfi, E. T. (2021). Analisis Perbandingan Akurasi Algoritma Naïve Bayes Dan C4.5 untuk Klasifikasi Diabetes. *Edumatic: Jurnal Pendidikan Informatika*, *5*(2), 147–156. https://doi.org/10.29408/edumatic.v5i2.3424

[3] Eko Aris Setiawan, & Nurhidayah, D. A. (2021). Universitas Muhammadiyah Ponorogo. *Edupedia*, *5*(2), 145–154. http://studentjournal.umpo.ac.id/index.php/edupedia

[4] Etriyanti, E., Syamsuar, D., & Kunang, N. (2020). Implementasi Data Mining Menggunakan Algoritme Naive Bayes Classifier dan C4.5 untuk Memprediksi Kelulusan Mahasiswa. *Telematika*, *13*(1), 56–67.

https://doi.org/10.35671/telematika.v13i1.881
[5] Fajri, F. N. (2018). *Perbandingan Sistem Klasifikasi Naïve Bayes dan Decision Tree Untuk Diagnosa Penyakit Diabetes*
[6] Fatmawati, F., & Narti, N. (2022). Perbandingan Algoritma C4.5 dan Naive Bayes Dalam Klasifikasi Tingkat Kepuasan Mahasiswa Terhadap Pembelajaran Daring. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *4*(1), 1–12. https://doi.org/10.35746/jtim.v4i1.196
[7] Genisa, L., & Mulyana, D. I. (2021). Implementasi Penerapan Metode C4.5 dan Naïve Bayes Dalam Tingkat Kelulusan Akreditasi Lembaga PAUD Pada Badan Akreditasi Nasional. *Jurnal Media Informatika Budidarma*, *5*(4), 1595. https://doi.org/10.30865/mib.v5i4.3267
[8] Handayani, K., Lisnawanty, L., Latif, A., Firdaus, M. R., & Hasan, F. N. (2021). Komparasi Algoritma C4.5 Dan Naïve Bayes Dalam Penentuan Status Kelayakan Donor Darah. *Sistemasi*, *10*(3), 676. https://doi.org/10.32520/stmsi.v10i3.1440
[9] Ibrahim, I., Gani, H., Lamusu, R., & Humolungo, Y. (2022). Perbandingan Algoritma Naïve Bayes Dan C4.5 Untuk Klasifikasi Bantuan Rumah Sehat. *Jurnal Ilmu Komputer (JUIK)*, *2*(1), 72. https://doi.org/10.31314/juik.v2i1.1477
[10] Sugara, B., Adidarma, D., & Budilaksono, S. (2019). Perbandingan Akurasi Algoritma C4.5 dan Naïve Bayes untuk Deteksi Dini Gangguan Autisme pada Anak. *Jurnal IKRA-ITH Informatika*, *3*(1), 119–128.
[11] Wardani, N. W., & Ariasih, N. K. (2019). Analisa Komparasi Algoritma Decision Tree C4.5 dan Naïve Bayes untuk Prediksi Churn Berdasarkan Kelas Pelanggan Retail. *International Journal of Natural Science and Engineering*, *3*(3), 103. https://doi.org/10.23887/ijnse.v3i3.23113
[12] Widayati, Y. T., Prihati, Y., & Widjaja, S. (2021). Analisis Dan Komparasi Algoritma Naïve Bayes. *Transformatika*, *18*(2), 161–172.
[13] Yulianti, I. (2019). Analisis Komparasi Klasifikasi Algoritma C4 . 5 Dan Naïve. *Jurnal Swabumi*, *7*(1), 62–66.