Volume 18 Nomor 1 (2023) 73-79
E – Link P-ISSN 1858-2109 E-ISSN 2656-5676
Jurnal Teknik Elektro dan Informatika

# KOMPARASI METODE KLASIFIKASI *DATA MINING* UNTUK PREDIKSI PENYAKIT STROKE

**Fitri Adha Hariyati Airi[1)] , Tati Suprapti[2)] , Agus Bahtiar[3)]**
[1,2,3)] Teknik Informatika, STMIK IKMI Cirebon, Kota Cirebon, Jawa Barat
E-mail: [1)] firiadha08@gmail.com, [2)] tatisuprapti112004@gmail.com, [3)] agusbahtiar038@gmail.com

**ABSTRAK**

Stroke merupakan penyakit dengan kondisi bahaya dan menjadi penyebab kematian nomor tiga setelah penyakit jantung koroner dan kanker. Kurangnya pengetahuan menjadikan masyarakat tidak menyadari tanda-tanda yang mungkin sudah terlihat. Apabila masyarakat mendapatkan pengenalan tentang penyakit stroke diharapkan dapat mengurangi dampak paling parah yaitu kematian. Oleh karena itu perlu dilakukan sebuah prediksi menggunakan metode klasifikasi. Hasil prediksi yang akurat dapat memudahkan para praktisi kesehatan dalam mengambil keputusan yang tepat. Data yang diambil merupakan data bersifat *public* dari situs *kaggle*. Pada penelitian ini *Orange* digunakan sebagai perangkat lunak. Penelitian ini melakukan sebuah perbandingan algoritma *Naive Bayes*, *K-Nearest Neighbor* dan *Random Forest*. Hasil yang diperoleh pada penelitian ini untuk algoritma *Naive Bayes* sebesar 71.9% *accuracy*, 71.7% *precision*, 71.9% *recall*. Sedangkan untuk algoritma *K-NN* mendapatkan nilai *accuracy* sebesar 73.6%, *precision* sebesar 73%, *recall* 73.6% dan untuk algoritma *Random Forest* mendapatkan nilai *accuracy* sebesar 92.5%, *precision* 92.5%, *recall* 92.5%.

**Kata kunci** : stroke, klasifikasi, algoritma

***ABSTRACT***

*Stroke is a dangerous disease and the third leading cause of death after coronary heart disease and cancer. Lack of knowledge makes people unaware of signs that may have been seen. If people get an introduction to stroke disease, it is hoped that it can reduce the most severe impact, namely death. Therefore, it is necessary to do a prediction using the classification method. The data taken is public data from the kaggle site. In this research Orange is used as software. This study conducted a comparison of Naive Bayes, K-Nearest Neighbor and Random Forest algorithms. The results obtained in this study for the Naive Bayes algorithm were 71.9% accuracy, 71.7% precision, 71.9% recall. Meanwhile, the K-NN algorithm gets an accuracy value of 73.6%, precision of 73%, recall of 73.6%. And, for the Random Forest algorithm, the accuracy value is 92.5%, precision 92.5%, recall 92.5%. Based on the results obtained, it proves that the Random Forest algorithm is better.*

***Keywords****: stroke, classification, algorithm*

## 1. PENDAHULUAN

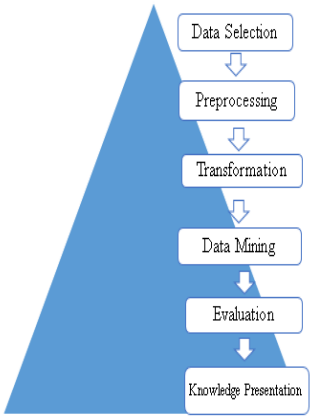
Penyakit yang menjadi persoalan didunia ini bisa menyerang semua kelompok umur, tetapi mayoritas stroke menyerang pada kelompok usia lanjut[1]. Stroke merupakan penyakit dengan kondisi bahaya dan menjadi penyebab kematian nomor tiga di dunia setelah penyakit jantung koroner dan kanker, karena berdasarkan data Organisasi Stroke Dunia setiap tahun terdapat 13,7 juta penderita stroke dan 5,5 juta kasus kematian akibat stroke[2]. Stroke menjadi salah satu masalah neurologik nomer satu yang ada di dunia. Indonesia menjadi negara di Asia dengan penderita stroke terbesar[3]. Stroke juga merupakan penyebab

paling tinggi pada gangguan fungsional dengan 20% penderita yang bertahan hidup membutuhkan perawatan institusi setelah 3 bulan dan 15-30% menjadi cacat permanen [4]. Kurangnya pengetahuan menjadikan masyarakat tidak menyadari tanda-tanda yang mungkin sudah terlihat, seperti : lengan yang sering merasa lemas, kesulitan berbicara, pusing dan kebingungan[5]. Maka dari itu, apabila masyarakat mendapatkan pengenalan tentang penyakit stroke sehingga penderita stroke dapat segera mencari pengobatan dengan tepat agar dapat mengurangi tingkat kecatatan maupun kematian [6]. Penanganan yang cepat dan tepat membantu penderita stroke dalam memulihkan penyakitnya, apabila penanganan terlambat dapat mengakibatkan kecacatan jangka panjang, kerusakan otak yang berkepanjangan, bahkan menyebabkan kematian[7]. Kasus ini mengundang banyak perhatian para peneliti untuk dapat memprediksi hasil yang akurat dengan harapan setelah adanya hasil prediksi ini dapat membantu para praktisi kesehatan agar dapat menangani dengan tepat sehingga mengurangi resiko terburuk yaitu kematian.. Penelitian ini menggunakan metode klasifikasi yang dimana metode ini banyak dikembangkan dengan bantuan komputasi cerdas yang mampu mengolah data dengan jumlah yang besar. Pada penelitian ini, beberapa metode klasifikasi yaitu *Naive Bayes*, *K-Nearest Neighbor* dan *Random Forest* diimplementasikan pada kasus penyakit stroke untuk kemudian dapat dibandingkan hasil *performance*-nya (Akurasi, Presisi, Recall).

Algoritma *Random Forest*, *Naive Bayes* dan *K-Nearest Neighbor* adalah algoritma pengklasifikasian *dataset* yang umumnya mendapatkan nilai akurasi relatif tinggi. Ketiga algoritma tersebut masing-masing memiliki kelebihan dan kekurangan. Dengan demikian, pada penelitian ini dilakukan perbandingan algoritma diatas untuk memperoleh algoritma yang paling cocok dalam klasifikasi terhadap penyakit stroke [8].

## 2. METODE PENELITIAN

Tahapan penelitian dilaksanakan seperti pada Gambar 1




**Gambar 1**. Tahapan Penelitian

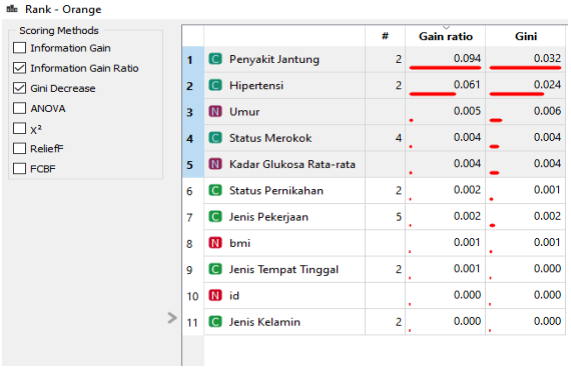
Tahapan yang akan dilakukan seperti pada Gambar 1 antara lain :

### *2.1. Data Selection*

Pada tahapan ini data publik dari situs kaggle akan diseleksi dengan cara melihat kesesuaian data dengan topik atau judul penelitian yang akan diteliti. Data yang dipakai pada penelitian ini berjumlah 1500 dari 5111 data dengan atribut id, jenis kelamin, umur, hipertensi, penyakit jantung, status pernikahan, jenis pekerjaan, tipe tempat tinggal, kadar glukosa rata-rata, indeks massa tubuh, status merokok dengan hasil ouput stroke.

### *2.2. Preprocessing*

Tahap *preprocessing* pada data ini tidak terjadi *missing value*, hal ini dapat melanjutkan tahapan selanjutnya yaitu pemilihan atribut untuk melakukan pengolahan data berdasarkan pertimbangan *Rank* yang ada.


Rank - Orange

Scoring Methods
- Information Gain
- Information Gain Ratio
- Gini Decrease
- ANOVA
- $\chi^2$
- ReliefF
- FCBF

| | | # | Gain ratio | Gini |
|---|---|---|---|---|
| 1 | Penyakit Jantung | 2 | 0.094 | 0.032 |
| 2 | Hipertensi | 2 | 0.061 | 0.024 |
| 3 | Umur | | 0.005 | 0.006 |
| 4 | Status Merokok | 4 | 0.004 | 0.004 |
| 5 | Kadar Glukosa Rata-rata | | 0.004 | 0.004 |
| 6 | Status Pernikahan | 2 | 0.002 | 0.001 |
| 7 | Jenis Pekerjaan | 5 | 0.002 | 0.002 |
| 8 | bmi | | 0.001 | 0.001 |
| 9 | Jenis Tempat Tinggal | 2 | 0.001 | 0.000 |
| 10 | id | | 0.000 | 0.000 |
| 11 | Jenis Kelamin | 2 | 0.000 | 0.000 |

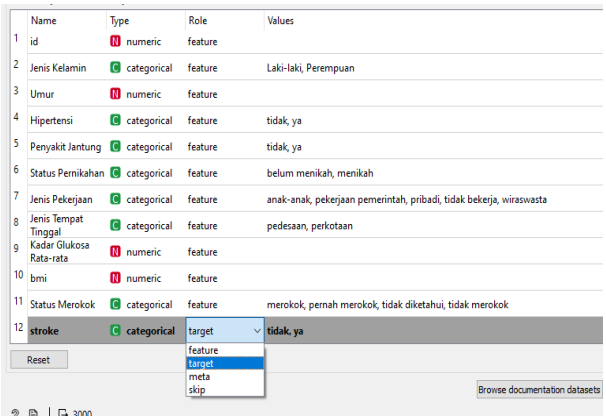
**Gambar 2.** Rank

Pada Gambar 2 menjelaskan bahwa penyakit jantung, hipertensi, umur, status merokok dan

kadar glukosa rata-rata menempati *Best Rank*, Karena hal itu lah atribut tersebut yang menjadi acuan untuk penelitian ini.

### *2.3. Transformation*

Setelah proses pembersihan data, selanjutnya dilakukan transformasi pada data sesuai dengan jenis data yang akan dikelompokkan menjadi dua kategori variabel. Penyakit jantung, hipertensi, umur, status merokok dan kadar glukosa rata-rata menjadi variabel prediktor dan stroke menjadi variabel target.




**Gambar 3**. Proses *Transformation* Data

### *2.4. Data Mining*

*Data mining* merupakan salah satu cara mengatasi masalah dengan analisa obyek yang telah ada dalam *data set*. Hasil dari *data mining* sendiri dapat dijadikan bantuan dalam pengambilan keputusan di masa mendatang. Pada langkah ini data mining yang diterapkan yaitu *Naive Bayes*, *K-Nearest Neighbor* dan *Random Forest*.

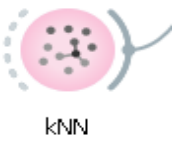
1. *Naive Bayes*

*Naive Bayes* merupakan salah satu metode *data mining* yang memanfaatkan perhitungan probabilitas. Algoritma ini juga sering digunakan untuk kebutuhan klasifikasi karena sifatnya yang sederhana [9]. Widget *Naive Bayes* pada aplikasi *Orange* dapat dilihat pada Gambar 4.


**Gambar 4.** Widget *Naive Bayes*

2. K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* merupakan salah satu cara guna menerapkan klasifikasi terhadap objek yang bersumber pada informasi dengan jumlah K yang sudah ditetapkan[10]. *Widget K-Nearest Neighbor* pada aplikasi *Orange* dapat dilihat pada Gambar 5.




**Gambar 5.** Widget *K-NN*

3. *Random Forest*

*Random Forest* digunakan untuk pengklasifikasian dataset dalam jumlah besar. Cara kerja algoritma ini yaitu dengan membangun beberapa pohon keputusan dan menggabungkannya untuk mendapatkan prediksi yang lebih baik [11]. *Widget Random Forest* pada aplikasi *Orange* dapat dilihat pada Gambar 6.


**Gambar 6.** Widget Random Forest

### *2.5. Evaluation*

Tahap evaluasi dilakukan setelah proses data yang dilakukan dengan algoritma Naive Bayes, *K-Nearest Neighbor* dan *Random Forest* menghasilkan berapa banyak yang terkena stroke dan tidak terkena stroke yang kemudian akan dihitung tingkat akurasi, presisi maupun recall

a. Rumus menghitung Accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (1)$$

b. Rumus menghitung Precision

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

c. Rumus menghitung Recall

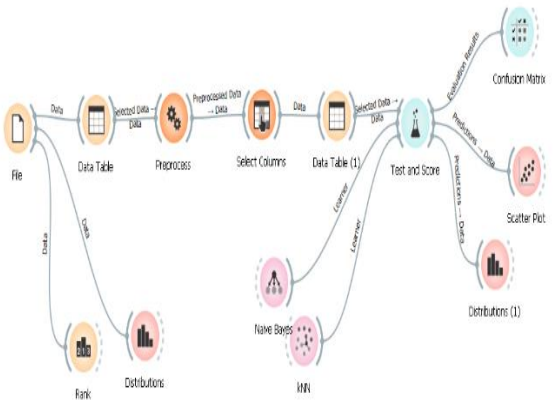
$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Ket :

- TP : *True Positive*
- TN : *True Negative*
- FP : *False Positive*
- FN : *False Negative*

### *2.6. Knowledge Presentation*

*Knowledge Presentation* menunjukkan proses utama yang dilakukan melalui aplikasi *Orange*. Proses utama tersebut dapat dilihat pada Gambar 7.




**Gambar 7.** Model Proses Klasifikasi

## 3. HASIL DAN DISKUSI

Hasil yang diperoleh dari proses penelitian, diantaranya :

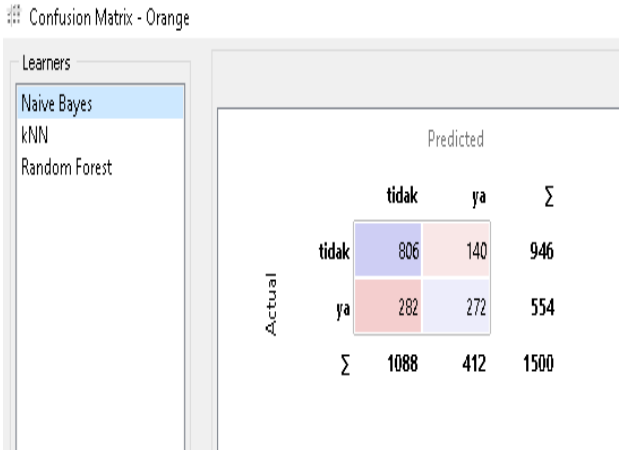
### *3.1 Confusion Matrix*

*Confusion matrix* merupakan salah satu cara untuk mengukur performa dari masing-masing algoritma.

Hasil *confusion matrix* dengan metode *Naive Bayes* seperti pada Gambar 8.

1. Dari penjelasan gambar 8 tentang hasil *confusion matrix*, dapat ditarik kesimpulan sebagai berikut :

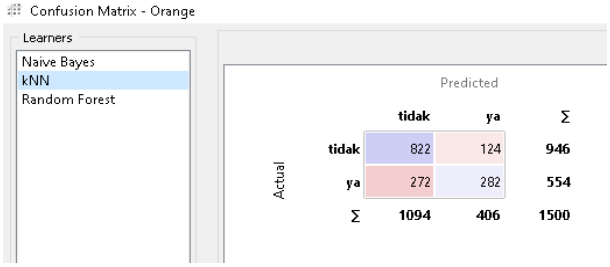
**Pertama**, untuk *True Negative* (TN) sebanyak 140, *True Positive* (TP) 806 dari data diprediksi tidak terkena stroke sebanyak 946.




**Gambar 8**. Hasil *Confusion Matrix* Metode *Naive Bayes*

**Kedua**, untuk *True Negative* (TN) sebanyak 282, *True Positive* (TP) 272 dari data yang terkena stroke sebanyak 554

2. Hasil confusion matrix dengan metode *K-Nearest Neighbor* seperti ditunjukkan pada Gambar 9.



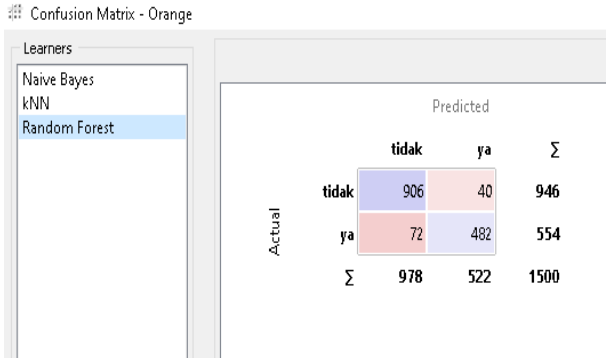

**Gambar 9**. Hasil Confusion Matrix Metode *K-NN*

Dari gambar 9 yang menjelaskan hasil *confusion matrix* dapat ditarik kesimpulan bahwa :

**Pertama**, untuk *True Negative* (TN) sebanyak 124, *True Positive* (TP) 882 dari data diprediksi tidak terkena stroke sebanyak 946.

**Kedua**, untuk *True Negative* (TN) sebanyak 272, *True Positive* (TP) 2282 dari data yang terkena stroke sebanyak 554

3. Hasil *confusion matrix* dengan metode *Random Forest* seperti pada Gambar 10.




**Gambar 10.** Hasil *Confusion Matrix* Metode *Random Forest*

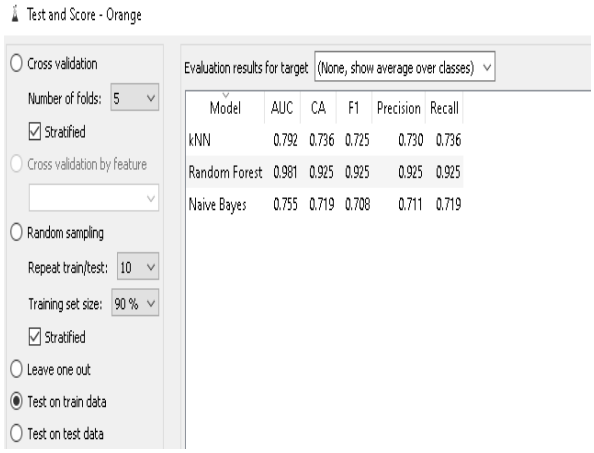
Dari gambar 10 tentang hasil *confusion matrix*, dapat ditarik kesimpulan sebagai berikut :

**Pertama**, untuk *True Negative* (TN) sebanyak 40, *True Positive* (TP) 906 dari data diprediksi tidak terkena stroke sebanyak 946.

**Kedua**, untuk *True Negative* (TN) sebanyak 72, *True Positive* (TP) 482 dari data yang terkena stroke sebanyak 554

### *3.2 Test and Score*

*Test and score* merupakan cara untuk melihat tingkat *Accuracy*, *Precision* dan *Recall* dari masing-masing algoritma yang sudah dikelola. Hasil dari *Test and Score* pada penelitian ini bisa dilihat pada Gambar 11.




**Gambar 11.** Hasil Akurasi dalam Bentuk Desimal

Pada gambar data penyakit stroke sebanyak 1500 setelah dilakukan proses pengujian, maka diperoleh hasil perhitungan *Accuracy*, *Presicion* dan *Recall* dengan menerapkan 90% dari data sebagai data latih dan 10% sebagai data uji melalui aplikasi orange setiap model seperti pada tabel 1.

**Tabel 1.** Hasil Akurasi Dalam Bentuk Desimal

| Metode | Accuracy | Precision | Recall |
|---|---|---|---|
| Naive Bayes | 0.719 | 0.711 | 0.719 |
| KNN | 0.736 | 0.730 | 0.736 |
| Random Forest | 0.925 | 0.925 | 0.925 |

Selanjutnya, hasil dari tabel akan diubah menjadi bentuk persen yang akan ditampilkan melalui tabel 2.

**Tabel 2.** Hasil Akurasi Dalam Bentuk Persen

| Metode | Accuracy | Precision | Recall |
|---|---|---|---|
| Naive Bayes | 71.9% | 71.1% | 71.9% |
| KNN | 73.6% | 73% | 73.6% |
| Random Forest | 92.5% | 92.5% | 92.5% |

Dari hasil perbandingan untuk model *Naive Bayes*, *K-NN* dan *Random Forest* maka diperoleh nilai akurasi tertinggi adalah metode *Random Forest* yaitu 92.5%

## 4. KESIMPULAN DAN SARAN

Berdasarkan hasil klasifikasi yang sudah dilakukan, diperoleh suatu kesimpulan dan saran, yaitu :

### 4.1 Kesimpulan

Dari penelitian ini setelah menggunakan metode *Naive Bayes*, *K-Nearest Neighbor* dan *Random Forest* untuk mengklasifikasi penyakit

stroke didapatkan hasil yang menyatakan bahwa algoritma *Random Forest* mendapatkan hasil yang lebih baik daripada *Naive Bayes* dan *K-NN*. Dengan menerapkan 90% dari data sebagai data latih dan 10% sebagai data uji memiliki nilai *accuracy* sebesar 71.9%, *precision* 71.1% dan 71.9% untuk *recall* pada algoritma *Naive Bayes*. Sedangkan, untuk *K-NN* mendapatkan nilai *accuracy* sebesar 73.6%, *precision* 73%, *recall* sebesar 73.6%. Dan yang terakhir, untuk *Random Forest* mendapatkan nilai *accuracy* sebesar 92.5%, *precision* 92.5% dan *recall* sebesar 92.5%.

### 4.2 Saran

Pada penelitian selanjutnya ada beberapa saran yang dapat disampaikan yaitu sebagai berikut :

1. Jumlah *dataset* yang digunakan diharapkan lebih banyak dari penelitian ini, sehingga dapat meningkatkan pemberian informasi mengenai penyakit stroke serta akurasi yang didapatkan lebih baik lagi dari penelitian ini.
2. Menggunakan metode lain, seperti : algoritma c.45, AdaBost, dll sehingga dapat melihat setiap akurasi yang didapatkan dari proses penelitian.

## 5. DAFTAR PUSTAKA

[1] A. Nur wahyuni, A. Faadilah, A. Nurani Asmara, A. Rahayu, dan A. Koswara, "Pengaruh Penyuluhan Kesehatan Tentang Penyakit Stroke Terahdap Tingkat Pengetahuan Keluarga," *Kolaborasi Jurnal Pengabdian Masyarakat*, vol. 1, no. 1, hlm. 42–51, Nov 2021, doi: 10.56359/kolaborasi.v1i1.5.

[2] M. N. Maskuri, K. Sukerti, and R. M. Herdian Bhakti, "Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke Stroke Desease Predict Using KNN Algorithm," *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 4, no. 1, 2022.

[3] A. Puspitawuri, E. Santoso, dan C. Dewi, "Diagnosis Tingkat Risiko Penyakit Stroke Menggunakan Metode K-Nearest Neighbor dan Naïve Bayes," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer E-Issn*, vol. 3, no. 4, hlm. 3319–3324, Apr 2019.

[4] H. S. Utami, L. Fitriana, dan L. D. Saputri, "Sosialisasi Deteksi Dini Stroke Dengan Pemeriksaan Radiologi Mrimra Brain," dalam *Prosiding Seminar Nasional Lppm Ump*, 2022, hlm. 76–80.

[5] Y. Azhar, A. Khoiriyah Firdausy, and P. J. Amelia, "SINTECH Journal | 191 Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," vol. 5, no. 2, 2022, [Online]. Available: https://doi.org/10.31598

[6] D. U. M. Rachmad, H. Oktavianto, dan M. Rahman, "Perbandingan Metode K-Nearest Neighbor Dan Gaussian Naive Bayes Untuk Klasifikasi Penyakit Stroke," *Jurnal Smart Teknologi*, vol. 3, no. 4, hlm. 405–412, 2022.

[7] F. Handayani, "Pengetahuan Tentang Stroke, Faktor Risiko, Tanda Peringatan, Respon Mencari Bantuan, Dan Tatalaksana Pada Pasien Stroke Iskemik Di Kota Semarang," *Jurnal Ilmu Keperawatan Medikal Bedah*, vol. 2, no. 2, pp. 1–51, 2019.

[8] A. Purnamawati, W. Nugroho, D. Putri, and W. F. Hidayat, "Deteksi Penyakit Daun pada Tanaman Padi Menggunakan Algoritma Decision Tree, Random Forest, Naïve Bayes, SVM dan KNN," *InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan*, vol. 5, no. 1, 2020, doi: 10.30743/infotekjar.v5i1.2934.

[9] D. Ulhaq, N. Suarna, and G. Dwilestari, "Klasifikasi Berita Kriminal Menggunakan Algoritma Naive Bayes Berbasis PSO," vol. 6, 2022.

[10] M. Deni Akbar and Y. Arie Wijaya, "Klasifikasi Motif Batik Jawa Menggunakan Algoritma K-Nearest Neighbors (Knn)," vol. 10, 2022, [Online]. Available: https://ejournal.stmikgici.ac.id/

[11] D. Sudrajat, A. I. Purnamasari, A. Rinaldi, D. A. Kurnia, and A. Bahtiar, "Klasifikasi Mutu Pembelajaran Hybrid berdasarkan Algoritma C.45, Random Forest dan Naïve Bayes dengan Optimasi Bootsrap Areggating (Bagging) pada masa COVID-19," *Jurnal Riset Komputer*, vol. 9, no. 6, pp. 2407–389, 2022, doi: 10.30865/jurikom.v9i6.5179.