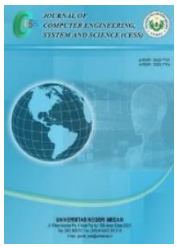
CESS
(Journal of Computing Engineering, System and Science) 7(1) January 2022 81-89

e-ISSN: 2502-714x
p-ISSN: 2502-7131

Contents list available at www.jurnal.unimed.ac.id

CESS
(Journal of Computing Engineering, System and Science)

journal homepage: https://jurnal.unimed.ac.id/2012/index.php/cess



Komparasi Akurasi Pada Naïve Bayes Dan Random Forest Dalam Klasifikasi Penyakit Liver

Comparison of Accuracy in Naïve Bayes and Random Forests in Classification of Liver Disease

Ahmadi Irmansyah Lubis[1*], Umri Erdiansyah[2], Rosma Siregar[3]

[1,2,3] STMIK TRIGUNA DHARMA
Jl. AH Nasution No 73F, 20142, Medan, Indonesia.
email: [1]ahmadi.loebis94@gmail.com, [2]umrierdiansyah13@gmail.com, [3]rosmasiregar@gmail.com,

Diterima: 31 Oktober 2021 | Diterima setelah perbaikan: 05 Nopember 2021 | Disetujui: 11 Desember 2021

A B S T R A K

Pada penelitian ini bertujuan untuk melakukan komparasi terhadap metode Naïve Bayes dan Random Forest dalam klasifikasi data pasien penyakit liver. Adapun data pengujian yang digunakan yaitu Indian Liver Patient Dataset (ILPD) yang diperoleh dari UCI iMachine Learning iRepository. Dataset tersebut memiliki 583 record data, 10 kriteria, dan 1 variable kelas serta dengan jumlah kelas sebanyak 2 kelas atribut, serta data set tersebut berjenis multivariate. Terdapat beberapa tahapan preprocessing yang dilakukan, antara lain normalisasi data yang diujikan, selanjutnya dilakukan analisis klasifikasi menggunakan metode naïvebayes dan random forest. Berdasarkan hasil pengujian yang dilakukan dalam memperoleh nilai akurasi perhitungan klasifikasi menggunakan Confusion Matrix, maka metode Random Forest memperoleh hasil yang terbaik yaitu dengan peroleh akurasi sebesar 70.60 % bila dibandingkan dengan Naïve Bayes yang hanya memperoleh akurasi sebesar 55.80 %. Sehingga Random Forest memiliki performa kinerja yang lebih unggul dalam perolehan akurasi yang dihasilkan dalam klasifikasi penyakit liver.

Kata Kunci: *Klasifikasi, Machine Learning, Naive Bayes, Random Forest, Liver.*

A B S T R A C T

This study aims to compare the Naïve Bayes and Random Forest methods in classifying liver disease patient data. The test data used is the Indian Liver Patient Dataset (ILPD) obtained from the UCI Machine Learning Repository. The dataset has 583 data records, 10 criteria, and 1 class variable and with a total of 2 attribute classes, and the data set is multivariate. There are several stages of preprocessing carried out, including normalization of the tested data, then classification analysis is carried out using the Naïve bayes method and random forest. Based on the results of the tests carried out in obtaining the accuracy of classification

**Penulis Korespondensi:*
email: ahmadi.loebis94@gmail.com

calculations using the Confusion Matrix, the Random Forest method obtained the best results, namely by obtaining an accuracy of 70.60% when compared to Naïve Bayes which only obtained an accuracy of 55.80%. So that Random Forest has a higher performance in terms of accuracy in the classification of liver disease.

Keywords: *Classification, Machine Learning, Naive Bayes, Random Forest, Liver.*

1. PENDAHULUAN

Salah satu penyakit yang terjadi pada hati manusia yang mengakibatkan peradangan dan merupakan salah satu unsur organ penting bagi tubuh manusia umumnya disebut sebagai Penyakit Liver. Adapun fungsi dari hati tersebut bagi tubuh manusia yaitu berfungsi dalam merubah zat beracun menjadi nutrisi serta dapat juga untuk mengendalikan hormon bagi tubuh manusia [1]. Menurut WHO (*World Health Organization*) tahun 2013, angka penderita penyakit liver di Indonesia mencapai 28 juta orang. Penyakit liver di indoneisa adalah merupakan salah satu dari 10 penyakit terbesar tingkat kemaitiannya [2].

Untuk mengidentifikasi penyakit liver yaitu melalui diagnosis sebagai proses untuk menentukan sifat suatu kelainan yang membedakannya dari keadaan yang mungkin terjadi. Mengindentifikasi penyakit liver dapat melakukan pemeriksaan fisik bagian pada tubuh termasuk paru-paru, jantung, kulit, otak, system saraf, dan perut yang dapat memberikan petunjuk untuk penyebab penyakit hati [3]. Dalam bidang ilmu komputer, proses mendiagnosis penyakit liver tersebut salah satunya dapat dilakukan dengan menerapkan proses klasifikasi.

Data Mining adalah proses menemukan pengetahuan menarik dari data dalam jumlah besar [4]. Adalah suatu proses ekstrasi atau penggalian data yang belum diketahui sebelumnya namun dapat dipahami dan berguna dari database yang besar serta digunakan untuk membuat suati keputusan bisnis yang sangat penting [5]. Data mining adalah ekstrasi informasi atau pola yang penting atau menarik dari data yang ada di database yang besar. Dan dalam jurnal data mining juga dikenal dengan *Knowleddge Discovery in Database* (KDD).

Ada beberapa penelitian terdahulu yang berkaitan dengan penelitian ini yaitu seperti pada tahun 2015, Rahmati melakukan penelitian dengan menggunakan Naïve bayes dan C4.5 pada data Pasien Penderita Liver. Untuk mengukur kinerja kedua algoritma tersebut digunakan metode pengujian *Cross Validation*, dan *Split Percentace*, dan pengukuranya dengan menggunakan *confusion matrix*. Hasil yang diperoleh yaitu C4.5 memiliki akurasi yang lebih tinggi dengan nilai 69.828% dibandingkan *Naïve Bayes* dengan nilai 63.362%. Dengan demikian C4.5 memberikan lebih baik untuk permasalahan mengidentifikasi penyakit Liver [6].

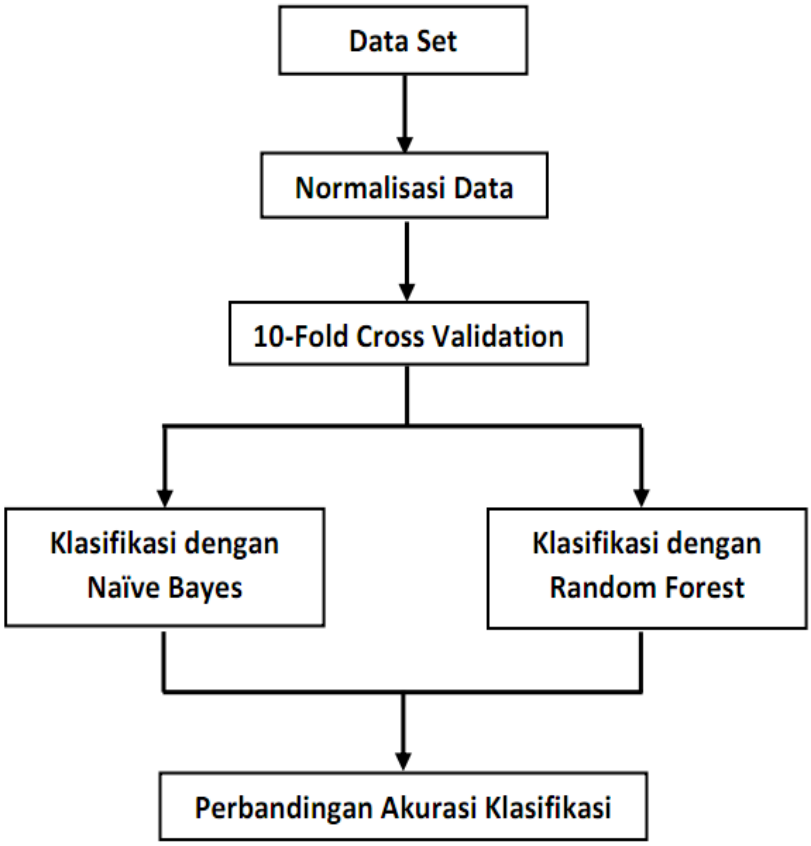
Kemudian pada tahun 2019, Pusporani *et al* melakukan penelitian dengan hasil penelitian tersebut yaitu SVM memberikan hasil yang terbaik secara akurasi, tapi berdasarkan *recall* K-NN memberikan hasil terbaik. Walaupun SVM memberikan hasil nilai akurasi dan presisi tertinggi tetapi terdapat ketimpangan yang besar antara nilai presisi dan recall yang dihasilkan, jika dibandingkan selisih nilai akurasi dan recall dari metode K-Nearest Neighbor [7].

Kemudian pada tahun 2019, Siburian & Mulyana melakukan penelitian dengan menerapkan *Random Forest* pada prediksi harga ponsel dengan tingkat akurasi prediksi yaitu sebesar 81%. Dan untuk nilai *Precision, Recall*, dan *F1-score* sebesar 81%. [8].

Maka berdasarkan penjabaran latar belakang pada penelitian ini, penulis akan menerapkan metode klasifikasi dalam mengindentifikasi penyakit liver menggunakan metode *Naïve Bayes* dan *Random Forest* yang kemudian melakukan komparasi dari kedua metode tersebut yang bertujuan untuk menentukan metode yang lebih akurat serta akurasi yang lebih baik dalam klasifikasi penyakit liver.

2. METODOLOGI PENELITIAN

Dalam menguji metode yang digunakan, dilakukan berdasarkan proses dari tahapan-tahapan penelitian pada Gambar 1 berikut:



Gambar 1. Tahapan Penelitian

Adapun penjelasan dari tahapan penelitian pada Gambar 1 yaitu sebagai berikut:

2.1. Dataset

Dataset yang digunakan yaitu bersumber dari *UCI Machine Learning Repository* dengan jenis Data yang digunakan pada penelitian ini berupa dataset *Indian Liver Patient Dataset* (ILPD) dengan jenis *Comma Separated Value* (csv). Dataset ini terdiri dari 583 *instance* dan 10 atribut dengan 1 label kelas yang bertipe teks yang terdiri dari dua nilai yaitu penderita liver dan bukan penderita liver. Adapun atribut dan jumlah kelas yang terdapat pada data set tersebut dapat dilihat pada Tabel 1 dan Tabel 2 berikut.

Tabel 1. Data Atribut Indian Liver Patient Dataset

Atribut	Penanda
Age	X1
Gender	X2
TB	X3
DB	X4
Alkphos	X5
SGPT	X6
SGOT	X7
TP	X8
ALB	X9
A/G	X10

Tabel 2. Kelas Atribut Indian Liver Pateint Dataset

Kelas	Penanda
Liver	1
Non-Liver	0

2.2. Normalisasi Data Set

Normalisasi data berfungsi untuk mempersiapkan data yang benar-benar valid sebelum diproses pada tahap berikutnya. Pada penelitian ini, normalisasi data dilakukan menggunakan metode *Min-Max* dengan rumus berikut.

$$\frac{(Data-Min)*(NewMax-NewMin)}{(Max-Min)} + NewMin \qquad (1)$$

2.3. 10-Fold Cross Validation

10-Fold Cross Validation untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua segmen, model pertama adalah untuk model *training* dan yang kedua ada untuk memvalidasi model *testing*. *10-Fold Cross Validation* membagi data set menjadi dua bagian yaitu *training data* dan *testing data* dengan proporsi, 90 % data latih, dan 10 % data uji. Cara kerja dari *10-fold cross validation* yaitu sebagai berikut:

a. Total *instance* dibagi menjadi *n* bagian.
b. *Fold* ke-1 adalah ketika bagian ke-1 menjadi data uji dan sisanya menjadi data latih dihitung menggunakan persamaan sebagai berikut.

$$\text{Akurasi} = \frac{\sum data\ uji\ benar\ klasifikasi}{\sum total\ data\ uji} 100\ X \qquad (2)$$

c. *Fold* ke-2 adalah ketika bagian ke-2 menjadi data uji dan sisanya menjadi data latih. Selanjutnya, menghitung akurasi berdasarkan porsi data tersebut menggunakan persamaan (2).

d. Demikian seterusnya hingga mencapai *fold* ke-k. Hitung rata-rata akurasi dari *k* buah akurasi di atas. Dan kemudian, rata-rata akurasi pada bagian ini menjadi akurasi final dari keseluruhan jumlah *k* yang digunakan dari awal pengujian sampai pada akhir pengujian.

2.4. Naïve Bayes

Naive Bayes termasuk salah satu algoritma dalam penggolongan probabilitas sederhana yang berdasarkan pada penerapan teori Bayes dengan asumsi indepedensi yang kuat (*naive*) pada fitur-fitur yang ada serta salah satu teknik sederhana untuk membangun klasifikasi model dengan menetapkan kelas untuk suatu masalah [9] [10].

Secara umum *Naive Bayes Classifier* memiliki kinerja yang baik jika dibandingkan dengan pengklasifikasian lain karena kesederhanaannya, kompleksitas waktu yang lebih sedikit, kebutuhan memori yang kecil, dan akurasi prediksi yang baik [11]. *Naive Bayes Classifier* dapat dinyatakan dalam persamaan:

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)} \qquad (3)$$

Keterangan:
C_i = Hipotesis data *X*
X = Data dengan kelas yang belum diketahui.
$P(C_i|X)$ = Probabilitas C_i berdasarkan kondisi *X*.
$P(X|C_i)$ = Probabilitas *X* berdasarkan kondisi C_i.
$P(C_i)$ = Probabilitas hipotesis C_i.
P(X) = Probabilitas dari data *X*.

2.5. Random Forest

Random Forest merupakan metode klasifikasi yang dalam proses kerjanya yaitu membangkitkan simpul anak untuk setiap *node* yang dilakukan secara acak untuk membangun pohon keputusan yang terdiri dari *root node*, *internal node*, dan *leaf node* dengan mengambil atribut dan data secara acak sesuai ketentuan yang diberlakukan. *Random Forest* dimulai dengan cara menghitung *entropy* sebagai penentu tingkat ketidakmurnian atribut dan *information gain*. Untuk menghitung *entropy* digunakan rumus seperti pada persamaan 4, sedangkan *information gain* menggunakan persamaan 5 berikut [12].

$$\boldsymbol{Entropy}(\boldsymbol{Y}) = -\sum_i \mathbf{p}(\mathbf{c}|\mathbf{Y})\ \mathbf{log_2}\ \mathbf{p}(\mathbf{c}|\mathbf{Y}) \qquad (4)$$

$$\boldsymbol{Information\ Gain}\ (\boldsymbol{Y}, \boldsymbol{a}) = \boldsymbol{Entropy}\ (\boldsymbol{Y}) - \sum \boldsymbol{veValues}\ ||\_\_\mathbf{YvYa}||\ \boldsymbol{Entropy}\ (\boldsymbol{Yv}) \quad (5)$$

2.6. Confusion Matrix

Confusion Matrix digunakan untuk menganalisis seberapa baik classifier mengenali data kelas yang berbeda [13]. Tabel tentang *Confusion Matrix* dapat dilihat pada tabel 3.

Tabel 3. Tabel Confusion Matrix

Actual Class	***Assigned Class***	
	Positive	***Negative***
Positive	*True Positive*	*False Negative*
Negative	*False Positive*	*True Negative*

True Positive dan *True Negative* adalah keadaan pada saat hasil prediksi sesuai dengan kondisi sebenarnya yang terjadi. *False Positive* dan *False Negative* adalah keadaan dimana hasil prediksi tidak sesuai dengan kondisi yang sebenarnya terjadi. Nilai akurasi *classifier* dapat dihitung persentase dari data uji yang diklasifikasikan dengan benar. Pengukuran presisi dan *recall* juga banyak digunakan dalam klasifikasi. Akurasi, presisi, dan *recall* dapat dihitung menggunakan rumus:

$$Acuraccy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$Precision = \frac{TP+TN}{TP+FP} \quad (8)$$

3. HASIL DAN PEMBAHASAN

3.1. Normalisasi Data Set

Dalam menormalisasi data yang akan diujikan, penulis menggunakan bantuan dari *RapidMiner Studio* untuk memudahkan proses normalisasi pada *Indian Liver Patient Dataset* (ILPD). Adapun hasil normalisasi data yang diperoleh dari perhitungan normalisasi *min-max* pada data set sebelum dan sesudah dinormalisasikan dapat dilihat pada Tabel 3 dan Tabel 4 berikut.

Tabel 4. Rincian Data ILPD Sebelum Dinormalisasi

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Class
65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
58	Male	1	0.4	182	14	20	6.8	3.4	1	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
26	Female	0.9	0.2	154	16	12	7	3.5	1	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	2
55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
.										
38	Male	1	0.3	216	21	24	7.3	4.4	1.5	2

Tabel 5. Hasil Normalisasi Min-Max Pada ILPD

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
0.709	1	0.004	0	0.060	0.003	0.001	0.594	0.521	0.24
0.674	0	0.140	0.275	0.310	0.027	0.018	0.695	0.5	0.176
0.674	0	0.092	0.204	0.208	0.025	0.011	0.623	0.521	0.236
0.627	0	0.008	0.015	0.058	0.002	0.002	0.594	0.543	0.28
0.790	0	0.046	0.096	0.064	0.008	0.009	0.666	0.326	0.04
0.488	0	0.018	0.030	0.070	0.004	0.001	0.710	0.760	0.4
0.255	1	0.006	0.005	0.044	0.003	0.004	0.623	0.565	0.28
0.290	1	0.006	0.010	0.067	0.002	0.002	0.579	0.586	0.32
0.151	0	0.006	0.010	0.067	0.006	0.001	0.681	0.695	0.36
0.593	0	0.004	0.005	0.110	0.021	0.009	0.594	0.543	0.28
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.395	0	0.008	0.010	0.074	0.005	0.002	0.666	0.760	0.48

3.2. Hasil Klasifikasi dengan Naïve Bayes

Setelah proses normalisasi data selesai dilakukan, maka selanjutnya dilakukan proses perhitungan klasifikasi. Tahapan klasifikasi yang pertama dilakukan yaitu dengan proses perhitungan klasifikasi pada *Indian Liver Patient Dataset* (ILPD) dengan perhitungan metode *Naïve Bayes.*

Untuk memudahkan proses klasifikasi, penulis dibantu dengan bantuan software *Waikato Environment for Knowledge Analysis* (WEKA). Adapun hasil klasifikasi yang diperoleh dengan Naïve Bayes yaitu pada Tabel 6 sebagai berikut.

Tabel 6. Hasil Klasifikasi Dengan Naïve Bayes

Fold Ke-	Jumlah Data Benar	Hasil Prediksi (%)		
		Accuracy	Precision	Recall
1	325	55.75	79.60	55.70
2	333	57.12	77.80	57.10
3	327	56.09	78.90	56.10
4	324	55.57	79.20	55.60
5	327	56.09	79.30	56.10
6	322	55.23	79.10	55.20
7	323	55.40	79.10	55.40
8	321	55.06	78.60	55.10
9	328	56.26	79.30	56.30
10	324	55.57	79.20	55.60
Rata-Rata		55.80	79.01	55.82

3.3. Hasil Klasifikasi dengan Random Forest

Kemudian dilakukan proses perhitungan klasifikasi pada *Indian Liver Patient Dataset* (ILPD) dengan perhitungan metode *Random Forest.* Adapun hasil klasifikasi yang diperoleh dengan *Random Forest* yaitu pada Tabel 7 sebagai berikut.

Tabel 7. Hasil Klasifikasi Dengan Random Forest

Fold Ke-	Jumlah Data Benar	Hasil Prediksi (%)		
		Accuracy	Precision	Recall
1	422	72.41	71.40	72.40
2	406	69.64	67.00	69.60
3	415	71.18	68.50	71.20
4	419	71.87	70.00	71.90
5	415	71.18	68.40	71.20
6	408	69.98	67.20	70.00
7	415	71.18	68.50	71.20
8	405	69.47	66.50	69.50
9	409	70.15	67.10	70.20
10	402	68.95	65.90	69.00
Rata-Rata		70.60	68.05	70.62

3.4. Komparasi Hasil Pengujian Klasifikasi

Setelah pengujian klasifikasi pada Naïve Bayes dan Random Forest selesai dilakukan, maka dapat dilihat hasil perbandingan dari hasil pengujian kedua metode tersebut dalam melakukan peroleh akurasi klasifikasi pada *Indian Liver Patient Dataset* (ILPD) yang dapat dilihat pada Tabel 8 berikut.

Tabel 8. Perbandingan Hasil Klasifikasi

Metode	Akurasi Prediksi (%)
Naïve Bayes	55.80
Random Forest	70.60

4. KESIMPULAN

Kesimpulan yang diperoleh setelah pengujian pada penelitian ini dengan menggunakan *Indian Liver Patient Dataset* (ILPD) diketahui dari hasil penelitian yaitu dengan metode Naïve Byaes diperoleh nilai akurasi 55.80 % dan metode *Random Forest* memperoleh nilai akurasi 70.60%. Sehingga dapat disimpulkan bahwa Naïve Bayes dan Random Forest memiliki performa kinerja yang berbeda di dalam proses perhitungan klasifikasi pada kasus penyakit liver yang dimana *Random Forest* lebih unggul peroleh akurasi yang dihasilkan.

REFERENSI

[1] M. Abdar, M. Zomorodi-Moghadam, R. Das, and I. H. Ting, "Performance analysis of classification algorithms on early detection of liver disease,". *Expert Systems with Applications*, vol. 67, pp. 239-251, 2017.

[2] P. M. C. Abrianto, "PENERAPAN METODE K-MEANS CLUSTERING UNTUK PENGELOMPOKKAN PASIEN PENYAKIT LIVER," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 2, no. 2, pp. 247-255, 2018.

[3] C. Y. Gobel, "Sistem Pakar Penyakit Liver Menggunakan K-Nearest Neighbors Algoritm Berbasis Website," *ILKOM Jurnal Ilmiah*, vol. 10, no. 2, pp. 152-159, 2018.

[4] J. Han, J. Pei, and M. Kamber, "Data mining: concepts and techniques," *Elsevier*, 2011.

[5] T. M. Connolly, and C. E. Begg, "Database systems: a practical approach to design, implementation, and management*," Pearson Education*, 2005

[6] E. Rahmawati, "Analisa Komparasi Algoritma Naive Bayes Dan C4. 5 Untuk Prediksi Penyakit Liver," *Jurnal Techno Nusa Mandiri*, vol. 12, no. 2, pp. 125-136, 2015.

[7] E. Pusporani, S. Qomariyah, and I. Irhamah, "Klasifikasi Pasien Penderita Penyakit Liver dengan Pendekatan Machine Learning," *Inferensi*, vol. 2, no. 1, pp. 25-32, 2019.

[8] V. W. Siburian, and I. E. Mulyana, "Prediksi Harga Ponsel Menggunakan Metode Random Forest," *In Annual Research Seminar (ARS)*, vol. 4, no. 1, pp. 144-147, 2019.

[9] M. R. Khan, S. K. Padhi, B. N. Sahu, and S. Behera, "Non stationary signal analysis and classification using FTT transform and Naive Bayes classifier," *2015 IEEE Power, Communication and Information Technology Conference, PCITC 2015 - Proceedings*, vol. 4, pp. 967–972, 2015.

[10] M. Granik, and V. Mesyura, "Fake news detection using naive Bayes classifier," *2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017 - Proceedings*, pp. 900–903, 2017.

[11] K. Netti, and Y. Radhika, "A novel method for minimizing loss of accuracy in Naive Bayes classifier," *2015 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2015*, pp. 1–4, 2015.

[12] M. Neshat, M. Sargolzaei, A. Nadjaran, and A. Masoumi, "Hepatitis disease diagnosis using hybrid casebased reasoning and particle swarm optimization," *International Scholarly Research Notices*, 2012.

[13] A. M. Siregar, and M. K. D. A. Puspabhuana, "Data Mining: Pengolahan Data Menjadi Informasi dengan RapidMiner," *CV Kekata Group,* 2017.