P-ISSN : 2549-7480
E-ISSN : 2549-6301

JURNAL ILMIAH INFORMATIKA
E-ISSN : 2549-6301
ISSN : 2549-7480
Faculty of Sains and Technology, Ibrahimy University
S4-Accredited : 105/E/KPT/2022
Journal page is available to : https://journal.ibrahimy.ac.id/index.php/JIMI

ANALISIS PERBANDINGAN PREDIKSI HARAPAN HIDUP HEPATITIS MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR DAN C4.5

Karina [1)], Herlina Hanum [2)], Anita Desiani [3)]

[1] *Program Studi Matematika, Universitas Sriwijaya*
[2] *Program Studi Matematika, Universitas Sriwijaya*
[3] *Program Studi Matematika, Universitas Sriwijaya*

email: [1] anita_desiani@unsri.ac.id, [2] linhanm@gmail.com, [3] 08011182025007@student.unsri.ac.id

ARTICLE INFO

Article History:
Recieved : 18 October 2023
Accepted : 20 December 2023
Published : 31 December 2023

Keywords:
Algorithm
C4.5
Hepatitis
K-Nearest Neighbor
Predictions

IEEE style in citing this article:
K. Karina, H. Hanum, and A. Desiani, "Analisis Perbandingan Prediksi Harapan Hidup Hepatitis Menggunakan Algoritma K-Nearest Neighbor dan C4.5", *Jurnal.ilmiah.informatika*, vol. 8, no. 2, pp. 98-111, Dec. 2023.

Corresponding Author:
Anita Desiani
Universitas Sriwijaya

ABSTRACT

Hepatitis is an inflammatory disease of the liver caused by a virus that causes damage to the cells and function of the liver. This study compares the accuracy, precision, and recall results of the K-Nearest Neighbor (K-NN) and C4.5 algorithms using the Percentage Split and K-fold Cross Validation methods. Of the two algorithms, the best level of accuracy is obtained using the K-fold Cross Validation method. Based on the accuracy and error rate, the best algorithm for predicting life expectancy for hepatitis sufferers is the K-NN algorithm. Based on the special Precision and Recall values on the Recall value to predict class zero the best algorithm is obtained using the C4.5 algorithm. To assess Precision and Recall, the other best algorithm in predicting the fixed response variable is obtained by using the K-NN algorithm. Overall, the best algorithm for predicting life expectancy for hepatitis sufferers is the K-Nearest Neighbor (K-NN) algorithm.

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

P-ISSN: 2549-7480
K. Karina dkk/ JIMI 8 (2) pp. 98-111 E-ISSN: 2549-6301

1. PENDAHULUAN

Hepatitis adalah suatu penyakit peradangan hati yang umumnya disebabkan oleh virus yang menyerang dan menyebabkan kerusakan pada sel-sel dan fungsi organ hati [1]. Penyakit ini banyak diderita baik orang dewasa maupun anak-anak [1]. Berdasarkan data riset kesehatan dasar (Riskesdas) tahun 2013 diperkirakan terdapat 28 juta penduduk yang terinfeksi hepatitis, 14 juta orang di antaranya berpotensi menjadi penderita hepatitis kronik dan 1,4 juta dari yang kronik tersebut berpotensi terkena kanker hati [2]. Bahkan apabila penderita hepatitis sudah kronik maka dapat menyebabkan kematian, sementara dokter tidak dapat menentukan harapan hidup penderita hepatitis [2]. Seiring adanya kemajuan di bidang teknologi, kini sudah dapat dimanfaatkan dalam bidang informatika salah satunya adalah teknologi data mining [3]. Data mining adalah serangkaian proses untuk menggali nilai dalam bentuk pengetahuan yang sebelumnya tidak diketahui secara manual dari suatu kumpulan data [4]. Dalam penelitian ini digunakan dua algoritma untuk membandingkan algoritma mana yang terbaik. Algoritma yang digunakan yaitu K-*Nearest Neighbor* (K-NN) dan C4.5. K-NN adalah suatu metode untuk mengklasifikasikan objek berdasarkan data training yang jaraknya paling dekat dengan objek tersebut [5]. C4.5 merupakan algoritma yang digunakan untuk membangun sebuah pohon keputusan (*decision tree*) dari data yang secara rekursif mengunjungi tiap simpul keputusan, memilih percabangan optimal, sampai tidak ada cabang lagi yang mungkin dihasilkan [6]. Pohon keputusan sendiri merupakan metode klasifikasi dan prediksi yang berguna untuk mengeksplor data dan menemukan hubungan yang tersembunyi dari variabel atau atribut yang digunakan, dan sebuah variabel target yang biasa disebut kelas [7]. Dari dua algoritma yang akan digunakan terdapat kelebihan dan kekurangan masing-masing.

Algoritma K-NN memiliki kelebihan kecepatan pelatihan yang sangat cepat, sederhana, dan mudah dipelajari [8], tangguh terhadap training data yang *noisy* dan efektif apabila data training besar [9]. Kekurangan K-NN, yaitu pada pemilihan nilai K yang perlu mempertimbangkan ukuran data, apabila data training terlalu besar maka waktu komputasi akan menjadi tinggi sedangkan apabila ukuran data terlalu kecil maka dimensi data dan variasi jarak dalam tabel jarak antara data latih menjadi lebih kecil sehingga peluang suatu data uji dikenali masuk ke kelas lain menjadi lebih besar [10]. Sementara itu, pada algoritma C4.5 merupakan algoritma yang sudah sangat terkenal dan disukai karena kelebihannya yang dapat mengolah data kategori dan diskrit, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diinterpretasikan dan performanya merupakan salah satu yang tercepat dibandingkan dengan algoritma lain, dan dapat menunjukkan atribut mana yang paling berpengaruh [11]. Adapun kekurangan algoritma C4.5 adalah sulit membaca data berjumlah besar [12].

Penelitian dalam hal memprediksi harapan hidup penderita hepatitis telah dilakukan oleh Septiani [13] menggunakan algoritma C4.5 dengan metode K-*fold Cross Validation* menghasilkan akurasi sebesar 77.29%. Oktaviani dkk [3] melakukan komparasi tingkat akurasi salah satunya menggunakan algoritma C4.5 untuk mengklasifikasikan keberlangsungan hidup pasien hepatitis menghasilkan akurasi sebesar 80.6452%. Oktaviani dkk [3] melakukan komparasi tingkat akurasi salah satunya menggunakan algoritma C4.5 untuk mengklasifikasikan

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

P-ISSN : 2549-7480
K. Karina dkk/ JIMI 8 (2) pp. 98-111 E-ISSN : 2549-6301

keberlangsungan hidup pasien hepatitis menghasilkan akurasi sebesar 80.6452%. Sulastri dkk [14] melakukan analisis perbandingan klasifikasi prediksi penyakit hepatitis salah satunya menggunakan algoritma K-*Nearest Neighbor* (K-NN) menghasilkan akurasi sebesar 93%. Dari hasil akurasi yang diperoleh rata-rata berada di atas 75% yang berarti algoritma yang digunakan dapat memprediksi harapan hidup penderita hepatitis dengan baik. Sayangnya penelitian tersebut tidak membandingkan hasil algoritma K-NN dan C4.5. Selain itu, penelitian tersebut hanya menunjukkan hasil algoritma berdasarkan nilai akurasinya saja. Untuk itu dalam penelitian ini akan membandingkan hasil kinerja algoritma K-*Nearest Neighbor* (K-NN) dan C4.5 berdasarkan nilai presisi, recall, dan akurasi. Hasil dari penelitian ini akan menunjukkan algoritma terbaik untuk memprediksi harapan hidup penderita hepatitis.

2. METODE PENELITIAN

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah dataset dengan format .csv yang diperoleh dari situs Kaggle (https://www.kaggle.com/datset/Odebreaker619/hepatitis-data). Data tersebut berjumlah 155 baris yang berisi informasi berbagai gejala yang dialami oleh penderita hepatitis dan prediksi harapan hidupnya. Jumlah atribut yang terdapat dalam data ini adalah sebanyak 20 atribut, 1 diantaranya atribut target atau variabel respon. Variabel respon dalam data ini adalah atribut kelas. Atribut kelas terdiri dari dua prediksi yaitu kelas nol dan kelas satu. Kelas nol adalah variabel respon untuk penderita hepatitis yang diprediksi meninggal dan kelas satu adalah variabel respon untuk penderita hepatitis yang diprediksi tetap hidup. Sementara atribut lainnya terdiri dari beberapa tipe data antara lain tipe data kategori, numerik, dan boolean. Atribut yang bertipe data numerik antara lain yaitu *age*, *bilirubin*, *alk_phosphate*, *sgot*, *albumin*, dan *protime*. Atribut yang memiliki tipe data kategori adalah *sex*. Atribut ini memiliki dua nilai yaitu kelas nol sebagai jenis kelamin laki-laki dan kelas satu sebagai jenis kelamin perempuan. Atribut yang bertipe data boolean antara lain yaitu atribut *steroid*, *antivirals*, *fatigue*, *malaise*, *anorexia*, *liver_big*, *liver_firm*, *spleen_palpable*, *spiders*, *ascites*, *varices*, dan *histology*. Atribut ini memiliki dua nilai yaitu kelas nol yang berarti salah dan kelas satu yang berarti benar. Atribut dan informasi data yang digunakan pada penelitian ini disajikan pada tabel (1) [14].

Tabel 1. Atribut dan informasi data

Atribut	Keterangan Atribut		Data Kosong
Age	Usia penderita.	7-78	0
Sex	Jenis kelamin penderita.	0=laki-laki, 1=perempuan	0
Steroid	Senyawa organic lemak sterol tidak terhidrolisis.	0=salah, 1=benar	1
Antivirals	Obat yang merusak replikasi virus.	0= salah, 1= benar	0
Fatigue	Kelelahan pada saraf dan otot.	0= salah, 1= benar	1

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

Malaise	Lemas, lesu, letih, dan sakit.	0= salah, 1= benar	1
Anorexia	Gangguan makan yang ditandai rasa takut berlebihan bila berat badan bertambah dan gangguan persepsi pada bentuk tubuh	0= salah, 1= benar	1
Liver big	Penyakit hati akibat virus dan penggunaan alkohol.	0= salah, 1= benar	10
Liver firm			11
Spleen palpable	Kerusakan organ jaringan limfatik.	0= salah, 1= benar	5
Spiders	Sekumpulan pembuluh darah abnormal dekat permukaan kulit.	0= salah, 1= benar	5
Ascites	Penumpukan cairan di rongga perut.	0= salah, 1= benar	5
Varices	Pembuluh darah vena yang membengkak dan tampak dekat dari permukaan kulit.	0= salah, 1= benar	5
Bilirubin	Senyawa pigmen berwarna kuning yang merupakan produk katabolisme enzimatik biliverdin oleh biliverdin reductase.	0,3-8	6
Alk_phosphate	Untuk mengukur tingkat enzim fosfatase alkali dalam darah.	26-295	29
Sgot	Enzim yang biasanya ditemukan pada hati (liver), jantung, otot, ginjal, hingga otak.	14-648	4
Albumin	Protein utama yang terdapat dalam darah manusia yang diproduksi oleh organ hati.	2,1-6,4	16
Protime	Disintesis oleh hati dan merupakan prekursor tidak aktif dalam proses pembekuan.	0-100	67
Histology	Pemeriksaan contoh sampel jaringan pada penderita hepatitis.	0=salah, 1=benar	0
Kelas	Variabel respon.	0=meninggal, 1=hidup	0

2.2 Persiapan Data

Tahap persiapan data dilakukan agar dapat mengurangi kesalahan dan bias dalam data penelitian sehingga data siap digunakan pada saat pemodelan. Dalam data hepatitis ini jumlah variabel respon yang diprediksi sebagai kelas nol berjumlah 123 baris sedangkan variabel respon yang diprediksi sebagai kelas satu hanya berjumlah 32 baris saja. Untuk itu, dalam data hepatitis ini dilakukan sampel ulang sebanyak 60 kali. Sampel ulang ini dilakukan dengan mengambil sampel secara acak dari data yang diprediksi

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

P-ISSN : 2549-7480
K. Karina dkk/ JIMI 8 (2) pp. 98-111 E-ISSN : 2549-6301

meninggal. Sehingga total data training pada kelas nol menjadi 92 baris dan total seluruh data training menjadi 215 baris. Selanjutnya agar hasil pemodelan dapat bekerja dengan baik, untuk beberapa atribut yang memiliki data kosong perlu diisi dengan suatu nilai. Untuk atribut yang bertipe data numerik diisi dengan nilai mean sedangkan untuk atribut yang bertipe data kategori diisi dengan nilai modus.

2.3 Standarisasi Data

Standarisasi data dilakukan untuk atribut yang memiliki rentang nilai yang cukup jauh untuk menghindari salah satu atribut yang mendominasi. Pada penelitian ini metode standarisasi data menggunakan *Min-Max Normalization* dengan persamaan (1) berikut [15].

$$normalized(a) = \frac{minRange\ a + (a\ minValue)(maxRange - minRange)}{maxValue - minValue} \quad (1)$$

2.4 Percentage Split

Hasil klasifikasi akan dites menggunakan k% dari data tersebut, dimana k adalah proporsi dari dataset yang digunakan untuk data testing [16]. Pembagian data pada penelitian ini sebesar 60% untuk data training dan 40% untuk data testing.

2.5 K-Fold Cross Validation

K-*fold Cross Validation* merupakan pendekatan alternatif yang membuat penggunaan informasi yang tersedia lebih efisien dengan langkah-langkah sebagai berikut [17].

1) Secara acak bagi sampel menjadi K bagian yang sama.
2) Untuk bagian ke-K, cocokkan model dengan bagian data K-1 lainnya, dan gunakan model ini untuk menghitung prediksi untuk setiap pengamatan di bagian ke-K.
3) Ulangi langkah di atas untuk K=1, 2, …., K dan gabungkan himpunan prediksi K untuk membuat sampel penuh dari nilai aktual dan prediksi.
4) Gunakan nilai aktual dan prediksi untuk menghasilkan setiap ukuran *goodness of fit* yang diinginkan.

2.6 Penerapan Algoritma K-Nearest Neighbor (K-NN)

Algoritma K-NN mencari kesamaan kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama [2]. Berikut ini langkah-langkah menggunakan Algoritma K-NN [18].

1) Menentukan parameter K (jumlah tetangga paling dekat), Parameter K pada testing ditentukan berdasarkan nilai K optimum pada saat training.
2) Menghitung kuadrat jarak euclid (*euclidean distance*) masing-masing objek terhadap data sampel yang diberikan menggunakan rumus jarak euclid pada persamaan (2) [19].

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (2)$$

Keterangan :

$d(a,b)$: jarak euclid
a_i : data training ke-i
b_i : data testing ke-i
i : baris ke-i dari tabel
n : jumlah data training

3) Mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak euclid terkecil.
4) Mengumpulkan kategori Y (klasifikasi *nearest neighbor*).
5) Didapat hasil klasifikasi.

2.7 Penerapan Algoritma C4.5

Algoritma C4.5 menggunakan konsep *information gain* atau *entropy reduction* untuk memilih pembagian yang optimal [13]. Beberapa tahapan dalam membuat sebuah pohon keputusan dalam algoritma C4.5 adalah sebagai berikut [6] :

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

1) Mempersiapkan data training. Data training biasanya diambil dari data histori yang sudah pernah terjadi sebelumnya dan sudah dikelompokkan dalam kelas-kelas tertentu.
2) Menghitung akar pohon. Akar pohon akan diambil dari atribut yang akan dipilih, dengan cara menghitung nilai *gain* dari masing- masing atribut, nilai *gain* yang paling tinggi akan menjadi akar pertama. Sebelum menghitung nilai *gain* dari atribut, perlu dihitung dahulu nilai *entropy*. Nilai *entropy* dapat dihitung menggunakan persamaan (3) berikut :

$$Entropy(P) = \sum_{i=1}^{n} -p_i * log_{2p_i} \quad (3)$$

Keterangan:
P : Himpunan Peristiwa
n : Jumlah Partisi P
p_i : Proporsi dari P_i terhadap P

Kemudian setelah nilai *entropy* pada masing-masing atribut sudah diperoleh maka hitung nilai *gain* dengan menggunakan persamaan (4) berikut:

$$Gain(P, A) = Entropy(P) \sum_{i=1}^{n} \frac{|P_i|}{|P|} Entropy(P_i) \quad (4)$$

Keterangan :
P : Himpunan Peristiwa
A : Fitur n : Jumlah partisi atribut A
$|P_i|$: Jumlah kasus pada partisi ke-i
$|P|$: Jumlah peristiwa dalam P

3) Ulangi langkah hingga semua baris mendapat kelas yang sama.

2.8 Analisis Hasil

Dengan algoritma yang diterapkan pada data hepatitis ini diperoleh sebuah tabel yang menunjukkan kinerja dari sebuah model klasifikasi yang memiliki data jawaban benar yang disebut dengan *Confusion matrix*. *Confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya [20]. Pada pengukuran kinerja menggunakan *Confusion matrix*, terdapat empat istilah sebagai representasi hasil proses klasifikasi yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) [20]. Dalam data ini TP berarti jumlah kelas nol yang diprediksi sebagai kelas nol, TN berarti jumlah kelas satu yang diprediksi sebagai kelas nol, FP berarti jumlah kelas nol yang diprediksi sebagai kelas satu, dan FN berarti jumlah kelas satu yang diprediksi sebagai kelas nol.

Dari *Confusion matrix* kemudian dihitung *akurasi*, *Precision*, dan *Recall*. Akurasi klasifikasi menunjukkan performansi model klasifikasi secara keseluruhan, dimana semakin tinggi akurasi klasifikasi, maka semakin baik performansi model klasifikasi, atau sebaliknya [21]. Akurasi klasifikasi dihitung dengan menggunakan persamaan (5) berikut.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision merupakan tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem yang dapat dihitung menggunakan persamaan (6) berikut [22].

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall merupakan pengukuran untuk data dengan klasifikasi positif yang benar oleh sistem yang dapat dihitung menggunakan persamaan (7) berikut [22].

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

3. HASIL DAN PEMBAHASAN

Dari hasil pengolahan data yang dilakukan diperoleh 215 data pengamatan dengan atribut sebanyak 20. Atribut yang menjadi label klasifikasi adalah kelas. Metode latihan yang digunakan adalah *Percentage Split* dan K-*Fold Cross Validation*

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

P-ISSN : 2549-7480
K. Karina dkk/ JIMI 8 (2) pp. 98-111 E-ISSN : 2549-6301

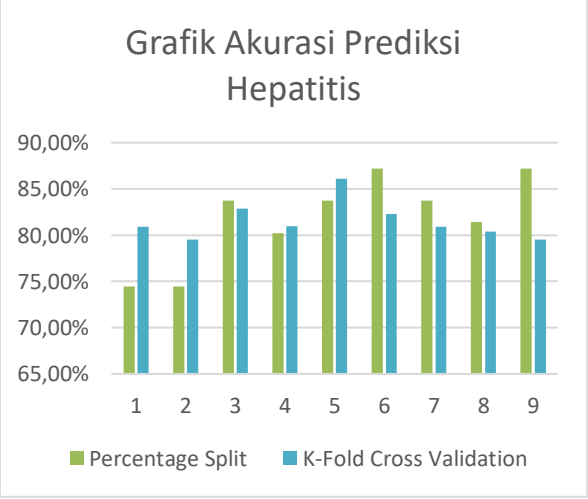
dengan data training 60% dan data testing 40%. Dari hasil ini akan didapatkan nilai akurasi, *Precision*, dan *Recall* untuk masing-masing algoritma.

3.1 Algoritma K-Nearest Neighbor (K-NN)

Dengan menggunakan algoritma K-NN percobaan dilakukan dengan 9 nilai K, yaitu 1, 2, 3, 4, 5, 6, 7, 8, dan 9 dengan metode *Percentage Split* dan *K-Fold Cross Validation*. Untuk perhitungan dengan jarak euclid diperoleh nilai *Precision*, *Recall*, dan akurasi seperti pada tabel (2). Kemudian nilai akurasi sistem dalam mendeteksi penyakit hepatitis berdasarkan penggunaan K pada sistem klasifikasi direpresentasikan pada gambar (1).

Tabel 1. Hasil pemodelan Algoritma K-NN

K	Target Kelas	Percentage Split			K-Fold Cross Validation		
		Precision	Recall	Accuracy	Precision	Recall	Accuracy
1	0	29%	25%	74.42%	77%	78%	80.91%
	1	83%	86%		84%	83%	
2	0	36%	50%	74.42%	71%	89%	79.52%
	1	88%	80%		90%	72%	
3	0	67%	35%	83.72%	80%	79%	82.86%
	1	85%	96%		85%	85%	
4	0	31%	33%	80.23%	74%	85%	80.95%
	1	89%	88%		87%	78%	
5	0	86%	32%	83.72%	86%	80%	86.08%
	1	84%	99%		86%	90%	
6	0	75%	40%	87.21%	78%	82%	82.30%
	1	88%	97%		86%	83%	
7	0	67%	13%	83.72%	81%	72%	80.93%
	1	84%	99%		81%	88%	
8	0	67%	22%	81.40%	78%	77%	80.41%
	1	82%	97%		83%	83%	
9	0	1%	15%	87.21%	79%	72%	79.52%
	1	87%	1%		80%	85%	



Gambar 1. Nilai akurasi hasil prediksi algoritma K-NN

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

Berdasarkan tabel 2 dan gambar 1 akurasi keseluruhan sistem, penggunaan K=5 memiliki akurasi keseluruhan yang terbesar dan lebih baik jika dibandingkan dengan K yang lain yaitu sebesar 83.72% pada *Percentage Split* dan 86.08% pada *K-Fold Cross Validation*. Dengan akurasi di atas 80% tersebut berarti algoritma K-NN dapat bekerja dengan baik dalam memprediksi harapan hidup penderita hepatitis. Dengan menggunakan metode *Percentage Split* diperoleh hasil pengukuran *Precision* untuk masing-masing kelas adalah 86% untuk kelas nol dan 84% untuk kelas satu dan hasil pengukuran *Recall* untuk masing-masing kelas adalah 32% untuk kelas nol dan 99% untuk kelas satu. Sementara dengan menggunakan metode *K-fold Cross Validation* diperoleh hasil pengukuran *Precision* untuk masing-masing kelas adalah 86% untuk kelas nol dan 86% untuk kelas satu dan hasil pengukuran *Recall* untuk masing-masing kelas adalah 80% untuk kelas nol dan 90% untuk kelas satu.

3.2 Algoritma C4.5

Hasil dari perhitungan gain pada algoritma C4.5 nilai gain tertinggi ada pada atribut albumin sehingga didapat bahwa atribut albumin adalah akar dari pohon keputusan. Nilai gain yang diperoleh menunjukkan jika penderita sudah memenuhi syarat nilai kadar albumin dalam darah penderita maka dapat diprediksi bahwa penderita hepatitis tersebut akan meninggal. Sebaliknya jika nilai kadar albumin dalam darah penderita belum mencukupi syarat maka harus melihat atribut lain untuk memprediksi harapan hidup penderita hepatitis. Dari perhitungan algoritma C4.5 diperoleh aturan linguistik untuk prediksi penderita hepatitis sebagai berikut:

IF albumin<= 3.85 **AND** age<= 28.5 **THEN** Prediksi Penderita Hepatitis Diprediksi= meninggal

IF albumin<= 3.85 **AND** age<= 28.5 **AND** sex<=0.963 **THEN** Prediksi Penderita Hepatitis= meninggal

IF albumin<= 3.85 **AND** age<= 28.5 **AND** sex<=0.963 **AND** bilirubin<= 1.3 **AND** age<= 48.0 **AND** histology<= 0.5 **AND** malaise<= 0.5 **THEN** Prediksi Penderita Hepatitis= die **OR** Prediksi Penderita Hepatitis= hidup

IF albumin<= 3.85 AND age<= 28.5 **AND** sex<=0.963 **AND** bilirubin<= 1.3 **AND** age<= 48.0 **AND** histology<= 0.5 **THEN** Prediksi Penderita Hepatitis= hidup

IF albumin<= 3.85 **AND** age<= 28.5 **AND** sex<=0.963 **AND** bilirubin<= 1.3 **AND** age<= 48.0 **AND** alk_phosphate <=0.336 **THEN** Prediksi Penderita Hepatitis= meninggal

IF albumin<= 3.85 **AND** age<= 28.5 **AND** sex<=0.963 **AND** bilirubin<= 1.3 **AND** age<= 48.0 **AND** alk_phosphate <=0.336 **AND** malaise <=0.5 **THEN** Prediksi Penderita Hepatitis= meninggal

IF albumin<= 3.85 **AND** age<= 28.5 **AND** sex<=0.963 **AND** bilirubin<= 1.3 **AND** age<= 48.0 **AND** alk_phosphate <=0.336 **AND** malaise <=0.5 **THEN** Prediksi Penderita Hepatitis=hidup

IF albumin<= 3.85 **AND** age<= 28.5 **AND** sex<=0.963 **AND** bilirubin<= 1.3 **AND** age<= 48.0 **AND** alk_phosphate <=0.972 **THEN** Prediksi Penderita Hepatitis= meninggal

IF albumin<= 3.85 **AND** age<= 28.5 **AND** sex<=0.963 **AND** bilirubin<= 1.3 **AND** age<= 48.0 **AND** alk_phosphate <=0.972 **AND** age<= 38.5 **THEN** Prediksi Penderita Hepatitis= hidup

IF albumin<= 3.85 **AND** age<= 28.5 **AND** sex<=0.963 **AND** bilirubin<= 1.3 **AND** age<= 48.0 **AND** alk_phosphate <=0.972 **AND** age<= 38.5 **OR** age<=36.5 **THEN** Prediksi Penderita Hepatitis= hidup

IF albumin<= 3.85 **AND** age<= 28.5 **AND** sex<=0.963 **AND** bilirubin<= 1.3 **AND** age<= 48.0 **AND** alk_phosphate <=0.972 **AND** age<= 38.5 **OR** age<= 36.5 **AND**

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

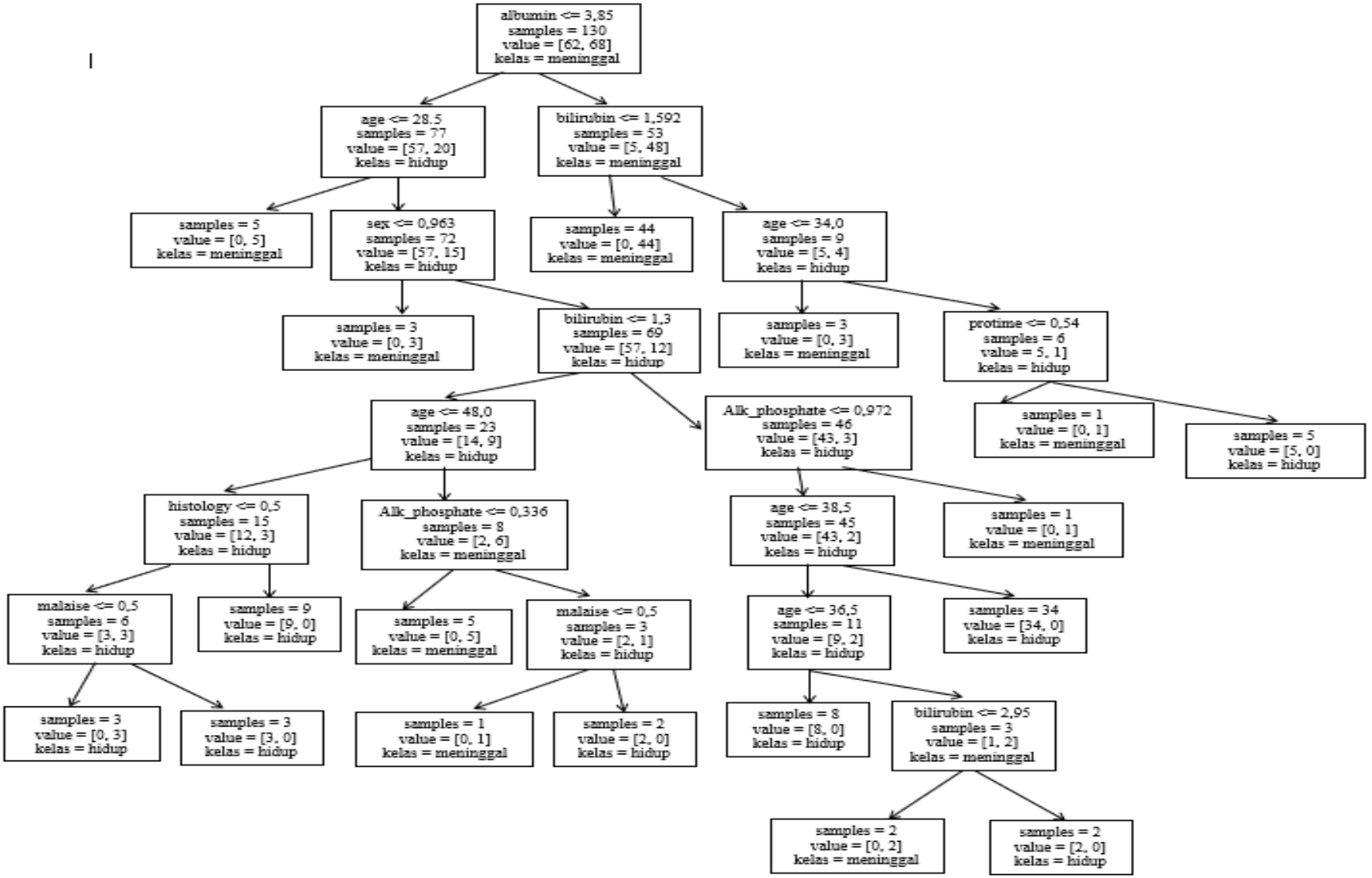
P-ISSN : 2549-7480
K. Karina dkk/ JIMI 8 (2) pp. 98-111 E-ISSN : 2549-6301

bilirubin<= 2.95 **THEN** Prediksi Penderita Hepatitis=die **OR** Prediksi Penderita Hepatitis= hidup
IF albumin<= 3.85 **AND** bilirubin<=1.592 **THEN** Prediksi Penderita Hepatitis= meninggal
IF albumin<= 3.85 **AND** bilirubin<=1.592 **AND** Age<= 34.0 **THEN** Prediksi Penderita Hepatitis= meninggal
IF albumin<= 3.85 **AND** bilirubin<=1.592 **AND** Age<= 34.0 **AND** protime<= 0.54 **THEN** Prediksi Penderita Hepatitis= die **OR** Prediksi Penderita Hepatitis= hidup

Secara lengkap aturan yang diperoleh dapat dilihat pada pohon keputusan gambar (2). Setelah didapatkan aturan keputusan dari pohon keputusan algoritma C4.5, proses uji mengukur sejauh mana keberhasilan model yang diperoleh dengan menggunakan *Confusion matrix*. Dengan menggunakan metode *Percentage Split* terdapat 24 data diprediksi secara benar sebagai kelas nol, 16 data yang seharusnya masuk dalam kelas nol tetapi diprediksi sebagai kelas satu. Kemudian terdapat 42 data diprediksi secara benar sebagai kelas satu, sedangkan 6 data yang seharusnya masuk dalam kelas satu tetapi diprediksi sebagai kelas nol. Sementara dengan menggunakan metode *K-fold Cross validation* terdapat 52 data diprediksi secara benar sebagai kelas nol, 19 data yang seharusnya masuk dalam kelas nol tetapi diprediksi sebagai kelas satu. Kemudian terdapat 49 data diprediksi secara benar sebagai kelas satu, sedangkan 10 data yang seharusnya masuk dalam kelas satu tetapi diprediksi sebagai kelas nol.

Selanjutnya dari *Confusion matrix* dapat dilakukan pengukuran tingkat akurasi yang diperoleh dari total data yang berhasil diprediksi secara benar sebesar 75% untuk metode *Percentage Split* dan 77,69% untuk metode K-*fold Cross Validation*. Hasil ini menunjukkan algoritma C4.5 cukup baik dalam memprediksi harapan hidup penderita hepatitis. Dengan menggunakan metode *Percentage Split* diperoleh hasil pengukuran *Precision* untuk masing-masing kelas adalah 60% untuk kelas nol dan 88% untuk kelas satu dan hasil pengukuran *Recall* untuk masing-masing kelas adalah 80% untuk kelas nol dan 72% untuk kelompok satu. Sementara dengan menggunakan metode *K-fold Cross Validation* diperoleh hasil pengukuran *Precision* untuk masing-masing kelas adalah 73% untuk kelas nol dan 83% untuk kelas satu dan hasil pengukuran *Recall* untuk masing-masing kelas adalah 84% untuk kelas nol dan 72% untuk kelas satu.

DOI : https://doi.org/10.35316/jimi.v8i2.98-111



Gambar 2. Pohon keputusan algoritma C.45

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

P-ISSN : 2549-7480
K. Karina dkk/ JIMI 8 (2) pp. 98-111 E-ISSN : 2549-6301

3.3 Perbandingan Hasil Kedua Algoritma

Hasil prediksi dua algoritma K-NN dan C4.5, menunjukkan bahwa baik dengan metode *Percentage Split* maupun dengan metode K-*fold Cross Validation* nilai akurasi prediksi pada algoritma K-NN selalu lebih tinggi dibandingkan pada algoritma C.45. Namun, meski akurasi prediksi algoritma K-NN selalu lebih tinggi, pada nilai *Precision* dan *Recall* tidak selalu nilai pada algoritma K-NN yang lebih tinggi. Dalam algoritma K-NN menggunakan metode *Percentage Split* data uji dikenali masuk ke kelas lain cukup besar karena tingkat ketepatan dalam memprediksi data yang benar cukup kecil yaitu hanya 32% saja. Selain itu, dalam algoritma K-NN juga tidak terlihat atribut mana yang paling mempengaruhi harapan hidup penderita hepatitis. Pada algoritma C4.5 jumlah data yang berhasil diprediksi secara benar melebihi dari setengah data yang ada sehingga data uji dikenali masuk ke kelas lain cukup rendah. Kemudian atribut yang saling mempengaruhi kemungkinan penderita hepatitis tetap hidup atau meninggal pun juga dapat dilihat dengan jelas. Dari hasil pohon keputusan dapat dilihat bahwa penderita hepatitis yang yang memiliki nilai kadar albumin dalam darah sesuai dengan standar nilainya, artinya penderita hepatitis akan meninggal sedangkan penderita hepatitis yang memiliki nilai kadar albumin dalam darah belum memenuhi syarat yang ada, artinya penderita hepatitis akan tetap hidup. Perbandingan hasil pengukuran kedua algoritma tersebut dapat dilihat pada tabel (3) .

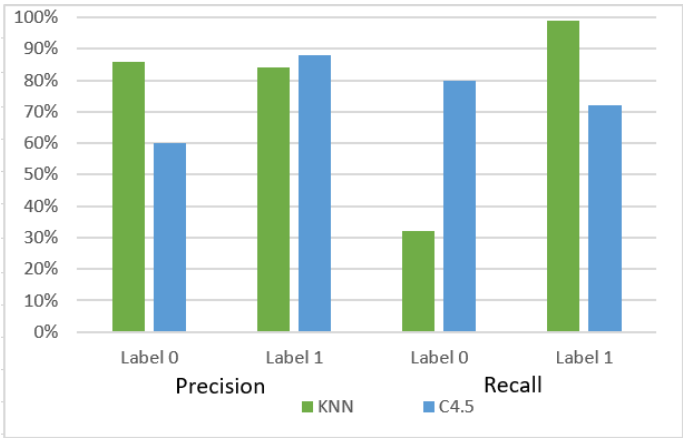
Tabel 3. Nilai Precision, Recall, dan Akurasi Algoritma K-NN dan C4.5

		Kelompok				
		Precision		Recall		Akurasi
Algoritma	Metode	0	1	0	1	
KNN	Percentage Split	86%	84%	32%	99%	83.72%
	K-fold Cross Validation	86%	86%	80%	90%	86.08%
C4.5	Percentage Split	60%	88%	80%	72%	75%
	K-fold Cross Validation	73%	83%	84%	72%	77.69%

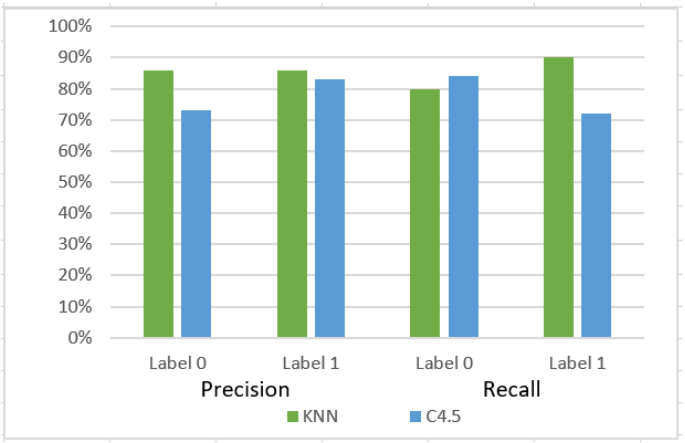
Hasil *Precision* dan *Recall* dari algoritma K-NN dan C4.5 dapat dilihat secara ringkas pada gambar (3) dan gambar (4). Dari gambar 3 dapat dilihat bahwa dengan metode *Percentage Split* nilai *Precision* pada kelas nol memang dapat diprediksi lebih baik menggunakan algoritma K-NN sementara pada kelas satu terlihat prediksinya sedikit lebih baik menggunakan algoritma C4.5. Begitupun dengan nilai *Recall* pada kelas nol dimana prediksi menggunakan algoritma C4.5 jauh lebih baik dibandingkan menggunakan algoritma K-NN, artinya meski akurasi dengan algoritma K-NN lebih tinggi namun, performanya dalam mengklasifikasikan kelas nol sangat kurang baik dibanding kelas satu sementara pada kelas satu prediksi yang lebih baik tetap diberikan oleh algoritma K-NN.

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

P-ISSN: 2549-7480
E-ISSN: 2549-6301



Gambar 3. Nilai *Precision* dan *Recall* menggunakan metode *Percentage Split* untuk Hasil Prediksi Hepatitis



Gambar 4. Nilai *Precision* dan *Recall* menggunakan metode *Percentage Split* untuk Hasil Prediksi Hepatitis

Dari gambar 4 dengan metode K-*fold Cross Validation* nilai *Precision* baik pada kelas nol maupun kelas satu algoritma yang dapat memprediksi lebih baik diberikan oleh algoritma K-NN. Sementara pada nilai *Recall* algoritma K-NN hanya dapat memprediksi lebih baik pada kelas nya performanya dalam memprediksi setiap kelas cukup baik.

4. KESIMPULAN

Hasil pengujian pada algoritma K-*Nearest Neighbor* (K-NN) menggunakan metode *Percentage Split* dan K-*fold Cross Validation* memberikan nilai akurasi, *Precision*, dan *Recall* yang terbaik dengan parameter K=1 sampai K=9, yaitu pada parameter K=5. Akurasi terbaiknya diperoleh dengan menggunakan metode K-*fold Cross Validation* yaitu di atas 85%. Begitupun dengan menggunakan algoritma C4.5 akurasi terbaik diperoleh dengan menggunakan metode K-*fold Cross Validation* yaitu di atas 75%. Artinya baik algoritma K-NN maupun C4.5 dapat bekerja dengan cukup baik dan paling baik menggunakan metode K- *fold Cross Validation*. Berdasarkan hasil perbandingan nilai akurasi prediksi terbaik ada pada implementasi algoritma K-NN. Berdasarkan nilai *Precision* satu saja sedangkan pada kelas nol dapat diprediksi lebih baik menggunakan algoritma C4.5 meskipun akurasinya berada di bawah algoritma K-NN. Pada nilai *Recall* ini baik dengan algoritma K-NN maupun C4.5 hasil prediksi pada setiap kelas berada di atas 70% yang arti-

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

P-ISSN : 2549-7480
K. Karina dkk/ JIMI 8 (2) pp. 98-111 E-ISSN : 2549-6301

algoritma terbaik dalam memprediksi kelas nol maupun kelas satu juga diperoleh dengan menggunakan algoritma K-NN sementara untuk nilai *Recall* algoritma terbaik dalam memprediksi kelas nol diperoleh dengan menggunakan algoritma C4.5. Secara keseluruhan algoritma terbaik dalam memprediksi harapan hidup penderita hepatitis ada pada algoritma K-*Nearest Neighbor* (K-NN).

5. REFERENSI

[1] J. L. Handarko and Alamsyah, "Implementasi Fuzzy Decision Tree Untuk Mendiagnosa Penyakit Hepatitis," *Unnes J. Math.*, vol. 4, no. 2, pp. 157–164, Nov. 2015.

[2] S. Khomsah, "Prediksi Harapan Hidup Penderita Hepatitis Kronik Menggunakan Metode-Metode Klasifikasi," *Semin. Nas. Inform. Medis.*, vol. 7, no. 1, pp. 38–45, Nov. 2018.

[3] P. S. Oktaviani, R. D. Ramadhani, T. G. Laksana, and A. E. Amalia, "Komparasi Tingkat Akurasi Support Vector Machine (SVM) dan C4.5 dalam Mengklasifikasikan Keberlangsungan Hidup Pasien Hepatitis," *Centive.*, vol 1, no. 1, pp. 163–167, Apr. 2019.

[4] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2016.

[5] M. Safaat, A. Sahari, and D. Lusiyanti, "Implementasi Metode K-Nearest Neighbor Untuk Mengklasifikasi Jenis Penyakit Katarak," *J. Ilm. Mat. Dan Terap.*, vol. 17, no. 1, pp. 92–99, Jun. 2020.

[6] M. W. Prihatmono and A. F. Watratan, "Implementasi Algoritma C4.5 Menggunakan Python Untuk Klasifikasi Kepuasan Konsumen," *Progres.*, vol. 11, no. 2, pp. 49–55, Sept. 2019.

[7] V. S. Ginting, K. Kusrini, and E. Taufiq, "Implementasi Algoritma C4.5 untuk Memprediksi Keterlambatan Pembayaran Sumbangan Pembangunan Pendidikan Sekolah Menggunakan Python," *Inspir. J. Teknol. Inf. dan Komun.*, vol. 10, no. 1, pp. 36–44, Jun. 2020.

[8] M. M. Siti Mutrofin, Abidatul Izzah, Arrie Kurniawardhani, "Optimasi Teknik Klasifikasi Modified K Nearest Neighbor Menggunakan Algoritma Genetika," *Gamma.*, vol. 10, no. 1, pp. 130–134, Sep. 2015.

[9] K. Eliyen, H. Tolle, and M. A. Muslim, "K-Nearest Neighbor Untuk Klasifikasi Penilaian Pada Virtual Patient Case," *J. Arus Elektro Indones.*, vol. 3, no. 1, pp. 13–18, Oct. 2017.

[10] A. Bode, "K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi Kopi Arabika," *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 188–195, Aug. 2017.

[11] F. F. Harryanto and S. Hansun, "Penerapan Algoritma C4.5 untuk Memprediksi Penerimaan Calon Pegawai Baru di PT WISE," *Jatisi.*, vol. 3, no. 2, pp. 95-103, Maret. 2017.

[12] L. M. N. Bernita, "Klasifikasi Persalinan Normal atau Caesar Menggunakan Algoritma C4.5," Fak. Sains dan Tekno. Universitas Sanata Dharma., Yogyakarta, Indonesia., Feb. 2017.

[13] W. D. Septiani, "Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis," *J. Pil. Nus. Mand.*, Volume, vol. 13, no. 1, pp. 76–84, Mar. 2017.

[14] S. Sulastri, K. Hadiono, and M. T. Anwar, "Analisis Perbandingan Klasifikasi Prediksi Penyakit Hepatitis Dengan Menggunakan Algoritma K-Nearest Neighbor, Naïve Bayes Dan Neural Network," *Dinamik*, vol. 24,

DOI : https://doi.org/10.35316/jimi.v8i2.98-111

no. 2, pp. 82–91, Jul. 2019.

[15] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, pp. 78-82, Jan. 2019.

[16] A. N. Kholifah and N. Insani, "Analisis Klasifikasi Pada Nasabah Kredit Koperasi X Menggunakan Decision Tree C4. 5 Dan Naive Bayes," *J. Pendidik. Mat. dan Sains*, vol. 5, no. 6, pp. 1–8, 2016.

[17] R. P. Ellis and P. G. Mookim, "K-Fold Cross-Validation is Superior to Split Sample Validation for Risk Adjustment Models," Department of Economics, Boston University., Boston, Amerika Serikat., 270 Bay State Road; Boston MA 02215, Jun. 2013.

[18] A. Y. Saputra and Y. Primadasa, "Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *Techno.Com*, vol. 17, no. 4, pp. 395–403, Nov. 2018.

[19] Rio Adi Arnomo, "Implementasi Algoritm. K-Nearest Neigbor untuk identifikasi Kualitas Air (Studi Kasus Pdam Kota Surakarta)," STMIK. Sinar Nusantara., Surakarta, Indonesia, Mar. 2017.

[20] Karsito and S. Susanti, "Klasifikasi Kelayakan Peserta Pengajuan Kredit Rumah Dengan Algoritma Naïve Bayes Di Perumahan Azzura Residencia," *J. Teknol. Pelita Bangsa*, vol. 9, no. 3, pp. 43–48, Mar. 2019.

[21] P. A. Octaviani, Yuciana Wilandari, and D. Ispriyanti, "Penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang," *J. Gaussian*, vol. 3, no. 4, pp. 811–820, 2014.

[22] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, Jul. 2020.

DOI : https://doi.org/10.35316/jimi.v8i2.98-111