JASIE "Jurnal Aplikasi Sistem Informasi dan Elektronika"
Vol. 3, No. 2, Desember 2021, Hal: 76-85
p-ISSN: 2714-612X, e-ISSN: 2798-1509

Pengaruh Teknik SMOTE Terhadap Prediksi Harapan Hidup Penderita Penyakit Hepatitis Menggunakan Algoritma *K-Nearest Neighbor*

Devi Putri Prianggi[1*], Agung Nilogiri[2], Reni Umilasari[3]
Program Studi Teknik Informatika, Universitas Muhammadiyah Jember[1,2,3]
Email: eviputri382@gmail.com[1*], agungnilogiri@unmuhjember.ac.id[2], reni.umilasari@unmuhjember.ac.id[3]

ABSTRAK

Hepatitis merupakan penyakit yang menyerang organ hati pada manusia, biasanya disebabkan oleh infeksi jamur, bakteri, penggunaan alkohol, autoimune, dan obat-obatan. Hepatitis sendiri terbagi menjadi 5, yaitu hepatitis A, B, C, D, dan E. Hepatitis B memiliki prevelensi tertinggi kedua di Asia Tenggara. Tujuan dari penelitian ini yaitu untuk mengetahui tingkat akurasi, presisi, dan *recall*. Dataset yang digunakan berasal dari Kaggle berjumlah 142 dan atribut sebanyak 20. Pada uji coba penelitian ini menggunakan *Cross Fold Validation* dengan nilai K = 2, 4, dan 5. Dari implementasi menggunakan algoritma *K-Nearest Neighbor* tanpa SMOTE diperoleh hasil validasi terbaik dengan akurasi tertinggi didapatkan 89,66% pada *Kfold* 5, presisi 78,12% pada *Kfold* 4, dan *recall* 94,44% pada *Kfold* 5. Sedangkan implementasi yang dilakukan menggunakan algoritma *K-Nearest Neighbor* menggunakan SMOTE diperoleh hasil validasi terbaik dengan akurasi tertinggi didapatkan 93,48 % pada *Kfold* 5, presisi 95,00 % pada *Kfold* 5, dan *recall* 92,11 % pada *Kfold* 5.

Kata Kunci: Prediksi Harapan Hidup, *K-Nearest Neighbor*, Teknik SMOTE, Hepatitis, *Cross Validation*

ABSTRACT

Hepatitis is a disease that attacks the liver in humans, usually caused by fungal infections, bacteria, alcohol use, autoimmune, and drugs. Hepatitis itself is divided into 5, namely hepatitis A, B, C, D, and E. Hepatitis B has the second highest prevalence in Southeast Asia. The purpose of this study is to determine the level of accuracy, precision, and recall. The dataset used comes from Kaggle totaling 142 and 20 attributes. In this research trial using Cross Fold Validation with values of K = 2, 4, and 5. From the implementation using the K-Nearest Neighbor algorithm without SMOTE, the best validation results were obtained with the highest accuracy obtained 89.66% on Kfold 5, 78.12% precision on Kfold 4, and 94.44% recall on Kfold 5. While the implementation carried out using the K-Nearest Neighbor algorithm using SMOTE obtained validation results the best with the highest accuracy obtained 93.48% on Kfold 5, 95.00% precision on Kfold 5, and recall 92.11% on Kfold 5.

Keywords: *Life Expectancy Prediction, K-Nearest Neighbor, SMOTE Technique, Hepatitis, Cross Validation*

1. PENDAHULUAN

Hati merupakan organ terpenting bagi manusia. Hati memiliki fungsi untuk membentuk dan mengeluarkan empedu, selain itu hati juga merupakan alat untuk penyimpanan glikogen, sintesis urea, metabolisme kolesterol dan lemak, serta detoksifikasi. Salah satu penyakit liver adalah hepatitis atau penyakit liver.

Hepatitis merupakan jenis penyakit endemik di beberapa negara berkembang, termasuk Indonesia. Ada lima jenis hepatitis, dari yang ringan hingga kronis, yaitu hepatitis A, hepatitis B, hepatitis C, hepatitis D, dan hepatitis E. Menurut data Riset Kesehatan Dasar (Riskesdas), pada tahun 2013 Indonesia memiliki prevalensi hepatitis B tertinggi kedua di Asia Tenggara. Hepatitis kronis, seperti hepatitis B, hepatitis C, dan hepatitis D, dapat berubah menjadi akut dan bisa menyebabkan sirosis dan kanker hati, pasien yang sudah dinyatakan mengidap hepatitis kronis bisa beresiko pada kematian (Khomsah, 2018). Hepatitis merupakan salah satu penyakit menular. Masyarakat Indonesia merupakan kelompok berisiko untuk tertular hepatitis A dan hepatitis E (Kemenkes RI, 2015).

Algoritma *K*-Nearest Neighbor merupakan algoritma untuk menentukan klasifikasi berdasarkan mayoritas dari *K*- tetangga terdekat (Ismail, 2018). Dari dataset atau data latih penelitian sering terjadi permasalahan-permasalahan diantaranya *imbalanced* data. Dataset dianggap tidak seimbang jika salah satu kelasnya memiliki dominan yang lebih besar dibandingkan kelas lainnya (Ali dkk, 2015).

Beberapa teknik digunakan untuk mengatasi ketidakseimbangan data. Salah satunya menggunakan *Synthetic Minority Oversampling Technique* (*SMOTE*). Teknik *SMOTE* adalah salah satu metode yang digunakan untuk menangani kasus ketidakseimbangan data dalam suatu dataset (Hidayah dkk. 2021).

Pada penelitian sebelumnya telah banyak digunakan metode untuk klasifikasi penyakit hepatitis dengan teknik data mining, diantaranya penelitian yang dilakukan oleh Septiani (2017) yaitu memprediksi penyakit hepatitis berdasarkan 155 data dan 2 atribut menggunakan metode C4.5 dan *Naïve Bayes*. C4.5 menghasilkan akurasi 77,29% dan nilai AUC 0,846 yang termasuk dalam Good Classification. Naive Bayes menghasilkan akurasi 83,71% dan nilai AUC 0,812. *Naïve Bayes* menghasilkan aurasi terbaik dibandingkan C4.5.

Penelitian yang dilakukan oleh Sulastri dkk. (2020) yaitu prediksi penyakit hepatitis berdasarkan 155 data dan 20 atribut menggunakan metode *K-Nearest Neighbor*, *Naïve Bayes* dan Neural Network. Algoritma *Naïve Bayes* tingkat akurasi yaitu 76.92%, tingkat error 23.01%, algoritma Neural Network tingkat akurasi yaitu 82,97%, tingkat error 17.03%, dan algoritma *K*-Nearest Neighbor tingkat akurasi yaitu 93%, tingkat error 7%. *K*-Nearest Neighbor termasuk tingkat akurasi terbaik dibandingkan *Naïve Bayes* dan Neural Network.

Berikutnya, Hidayah dkk. (2021) meneliti klasifikasi data pasien penderita gagal jantung berdasarkan 299 data menggunakan algoritma *K*-Nearest Neighbor menerapkan teknik SMOTE. Algoritma *K-Nearest Neighbor* tanpa SMOTE tingkat akurasi yaitu 71,59%, algoritma *K-Nearest Neighbor* menggunakan SMOTE tingkat akurasi yaitu 80,14. *K-Nearest Neighbor* menggunakan SMOTE tingkat akurasi terbaik dibandingkan *K-Nearest Neighbor* tanpa SMOTE, pada penelitian ini menggunakan dataset penyakit hepatitis yang di ambil dari KAGGLE sebanyak 142 data dan tools yang digunakan menggunakan phyton.

Berdasarkan kondisi kebutuhan penyakit hepatitis serta beberapa penelitian yang dijelaskan sebelumnya maka pada penelitian ini akan diterapkan Pengaruh teknik SMOTE terhadap prediksi harapan hidup penderita penyakit hepatitis menggunakan metode *K-Nearest Neighbor*.

2. KAJIAN PUSTAKA

A. Penyakit Hepatitis

Hepatitis merupakan masalah kesehatan global yang menyebabkan kematian pada bayi, anak kecil, dewasa dan lanjut usia. Kerentanan terhadap penularan penyakit hepatitis tidak menjadi masalah bagi berbagai pihak. Virus hepatitis mudah menular. Hepatitis adalah penyebab sirosis dan kanker hati, dan angka kematian akibat kanker tertinggi ketiga di dunia. Hepatitis menjadi ancaman bagi masyarakat karena kurangnya pengetahuan (Sari dkk, 2019).

Hepatitis penyakit yang menyerang hati manusia. Ini adalah hati, atau dimana hati menjadi meradang dan mempengaruhi fungsi hati. Ketika fungsi hati terganggu, maka fungsi organ lain juga ikut terganggu, sehingga merusak kesehatan seseorang secara keseluruhan. Akibat lainnya adalah hati menolak darah yang mengalir melaluinya menyebabkan tekanan darah tinggi dan pembuluh darah pecah. Gejala umum hepatitis termasuk sakit dan nyeri di sisi kanan perut, lemas, mual, demam, dan diare. Dalam beberapa kasus, gejala mirip flu dan penyakit kuning, yang membuat kulit dan mata terlihat kuning, juga terlihat. Namun, gejala hepatitis tidak selalu terlihat dan paling sering menyerang anak-anak. (Ramdhani, dkk., 2015).

Tabel 1. Atribut data

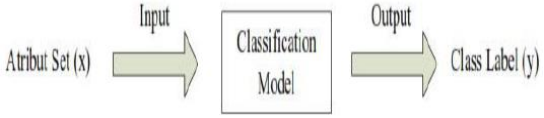
No	Atribut	Keterangan	
1	Age (Umur)	Angka numerik	Umur pasien
2	Sex (Jenis Kelamin)	1 = Laki-laki, 2 = Perempuan	Jenis kelamin pasien
3	Steroid	1 = No, 2 = Yes	Apa mendapatkan terapi steroid?
4	Antivirals	1 = No, 2 = Yes	Apa mendapatkan terapi antivirals?

5	Fatigue	1 = No, 2 = Yes	Apa mengalami gejala kelelahan akut
6	Malaise	1 = No, 2 = Yes	Apa mengalami gejala (rasa tidak nyaman)
7	Anorexia	1 = No, 2 = Yes	Apa mengalami gejala (muntah setiap makan)
8	Liver_big	1 = No, 2 = Yes	Apa kondisi hati/liver membesar?
9	Liver_firm	1 = No, 2 = Yes	Apa kondisi hati/liver mengeras?
10	Spleen_palpable	1 = No, 2 = Yes	Apa ada gejala limfa lebih jelas/besar dari normal?
11	Spiders	1 = No, 2 = Yes	Apa ada gejala pembuluh darah upnormal pada kulit (pembuluh darah mengumpul dan menonjol pada permukaan kulit)
12	Ascites	1 = No, 2 = Yes	Terjadi penumpukan cairan pada rongga perut
13	Varices	1 = No, 2 = Yes	Terjadi pembekakan vena esophagus (varices)
14	Bilirubin	0,39 - 4,00	Nilai kadar bilirubin dalam darah
15	Alk_phosphate	33 - 250	Kadar Alkalin Phostpate dalam liver
16	Sgot	13 - 500	Nilai SGOT
17	Albumin	2,1 - 6,0	Kadar albumin
18	Protime	10 - 90	Uji masa protombhine
19	Histology	1 = No, 2 = Yes	Apakah dilakukan pemeriksaan dengan (biopsy hati)
20	Class	1 = Die, 2 = Live	Label yang menunjukkan pasien hidup/mati

B. Klasifikasi

Klasifikasi ialah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui (Bustami, 2010).

Pada proses klasifikasi terbagi menjadi dua fase yaitu, latih dan uji. Pada fase latih, sebagian data yang telah diketahui kelas datanya (data latih) digunakan untuk membentuk model. Selanjutnya pada fase uji, model yang sudah dibentuk diuji dengan sebagian data lainnya (data uji) untuk mengetahui akurasi dari model tersebut (Han, 2006).



Gambar 1. Klasifikasi

C. Algoritma ***K***- *Nearest Neighbor*

K-Nearest Neighbor merupakan metode klasifikasi objek berdasarkan data pelatihan yang paling dekat dengan objek (Widiarsana dkk, 2011). Algoritma *K-Nearest Neighbor* metode yang digunakan untuk mengelompokkan objek berdasarkan sampel latih terdekatnya. Algoritma *K-Nearest Neighbor* bekerja dengan mencari kumpulan objek pada data training yang paling dekat dengan objek pada data test (Saxena, Khan, & Singh 2014). Rumus jarak euclidien sebagai berikut.

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

Keterangan:

D = Jarak kedekatan
X = Data testing
Y = Data training
N = Jumlah atribut 1 sampai n

D. *Synthetic Minority Oversampling Technique* (SMOTE)

Ketidakseimbangan data terjadi jika ada lebih banyak objek dalam satu lapisan data daripada di lapisan lainnya. Lapisan data dengan lebih banyak objek disebut kelas utama sedangkan lapisan lainnya disebut subkelas. Pengaruh penggunaan unbalanced data untuk membangun model sangat signifikan terhadap hasil model yang diperoleh (Hidayah dkk., 2021).

Metode SMOTE artinya, tambahkan data yang sedikit sehingga jumlahnya sama dengan data yang banyak. SMOTE menggunakan pendekatan lingkungan untuk menghasilkan data dari kelas kecil. Untuk data nominal diisi dengan nilai yang banyak dari k tetangga terdekat (Syukron dkk., 2020) dengan persamaan berikut.

$$X_{syn} = X_i + (X_{knn} - X_i)\ x\ \delta$$

Keterangan:

X_{syn} = data sintesis yang akan diciptakan
X_{knn} = data yang memiliki jarak terdekat dari data yang akan direplikasi
X_i = data dengan atribut ke-i
δ = nilai random antara 0 dan 1

E. *Confusion Matrix*

Confusion matrix adalah alat ukur yang digunakan untuk mengevaluasi keakuratan hasil klasifikasi dalam suatu penelitian. Matriks konfusi memiliki kolom yang berisi kelas fakta dan baris yang berisi kelas prediksi (Ting, 2017).

Confusion Matrix adalah alat yang berguna untuk menganalisis seberapa baik pengklasifikasi anda dapat mengenali data kelas yang berbeda. TP dan TN memberitahukan saat mengklasifikasi melakukan sesuatu dengan benar, sedangkan FP dan FN memberi tahu ketika klasifikasi melakukan kesalahan (Han dkk., 2012).

Tabel 2. *Confusion matrix* dua kelas

Kelas prediksi	Kelas aktual	
	Positif (+)	Negatif (-)
Positif (+)	TP	FP
Negatif (-)	FN	TN

Tabel 3. Rumus akurasi, presisi, dan *recall*

Akurasi	$\frac{TP + TN}{TP + TN + FP + FN}$
Presisi	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$

F. *Cross Validation*

Metode cross validasi yang digunakan sebagai mengevaluasi dan membandingkan pembelajaran dari 10 algoritma pembelajaran dengan membagi data menjadi dua bagian untuk pelatihan dan pengujian (Wibowo & Jumiati, 2018).

G. *Jupyter Notebook*

Jupyter ialah perangkat lunak yang bersifat terbuka *(open source)* dalam berbagai macam bahasa pemograman. Jupyter notebook ialah sebuah aplikasi web yang bersifat open source yang berfungsi untuk membuat dan membagikan dokumen berisi persamaan, visualisasi data dan lainnya (Widyatama & Suprapty, 2018). *Jupyter* merupakan salah satu perangkat lunak bahasa pemrograman

yang bersifat terbuka (open source) dan dirilis dibawah persyaratan liberal dan lisensi BSD yang dimodifikasi (Avila dkk, 2020).

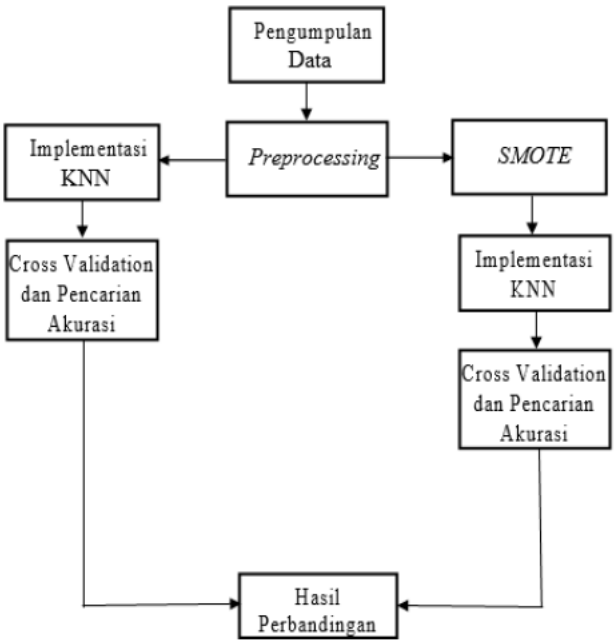
H. *Phyton*

Python adalah bahasa pemrograman yang serbaguna dan populer. Tidak seperti bahasa lain yang sulit dibaca dan dipahami, Python berfokus pada keterbacaan kode agar sintaks lebih mudah dipahami. Hal ini membuat Python sangat mudah dipelajari, baik untuk pemula maupun yang sudah mempelajari bahasa pemrograman lain. Bahasa tersebut pertama kali muncul pada tahun 1991 dan dirancang oleh seorang pria bernama Guido van Rossum (Hokya, 2013).

3. METODE PENELITIAN

A. Tahapan Penelitian

Dalam mengerjakan tugas akhir ini perlu adanya langkah-langkah penelitian yang mendukung dan maksimal dalam penyelesaian. Pada metode penelitian tentang pengaruh teknik SMOTE terhadap prediksi harapan hidup penderita penyakit hepatitis menggunakan algoritma *K-Nearest Neighbor*. Berikut adalah diagram metodologi penelitian yang memuat tahapan penelitian yang dilakukan sebagai berikut.



Gambar 2. Tahapan penelitian

B. Pengumpulan Data

Pengumpulan data yang dilakukan dalam penelitian ini diperoleh melalui Kaggle yang terdiri dari data penyakit hepatitis pada tahun 2019 yang dapat diakses melalui link berikut: https://www.kaggle.com/harinir/hepatitis. Dataset berupa jumlah kasus penyakit hepatitis pada tahun 2019 yang terdiri dari 142 data dengan 20 atribut antara lain, yaitu *age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver_big, liver_firm, spleen_palpable, spiders, ascites, varices, bilirubin, alk_phosphate, sgot, albumin, protime, histology, class*. Serta 2 kelas sebagai parameter output. Dalam penelitian ini data akan diolah menggunakan Algoritma *K-Nearest Neighbor* dan teknik SMOTE untuk mengetahui hasil akurasi, presisi, dan *recall*. Atribut yang digunakan berjumlah 20 dengan *output class* 1 dan 2, dimana 1 terklasifikasi meninggal dan 2 adalah hidup.

Tabel 4. Data latih

age	sex	steroid	…	histology	class
30	2	1	⋯	1	2
51	1	1	⋯	1	1
66	1	2	⋯	1	2
39	1	1	⋯	1	1
57	1	2	⋯	1	1
34	1	1	⋯	1	1
44	1	1	⋯	2	1
30	1	2	⋯	2	1
60	1	1	⋯	2	2
42	1	1	⋯	2	1
56	1	1	⋯	2	1
20	1	1	⋯	2	2
50	1	2	⋯	2	1
25	1	2	⋯	2	2
49	1	1	…	2	1

C. *Preprocessing*

Normalisasi data digunakan untuk menormalkan setiap atribut karena setiap atribut memiliki range data yang berbeda-beda. normalisasi data digunakan metode *MinMax Scaler* atau persamaannya adalah $X = \frac{X_i - X_{min}}{X_{max} - X_{min}}$, dimana X adalah hasil *scaler*, X_i adalah data ke i, X_{min} adalah data terkecil pada data keseluruhan dan X_{max} adalah nilai terbesar pada data keseluruhan.

D. Implementasi ***K****-Nearest Neighbor* Tanpa SMOTE

Tabel 5. Data latih

age	sex	steroid	…	histology	class
0,22	**1**	**0**	**…**	**0**	**2**
0,67	0	0	…	0	1
1,00	0	1	⋯	0	2
0,41	0	0	⋯	0	1
0,80	0	1	⋯	0	1
0,30	0	0	⋯	0	1
0,52	0	0	⋯	1	1
0,22	0	1	⋯	1	1
0,87	0	0	⋯	1	2
0,48	0	0	⋯	1	1
0,78	0	0	⋯	1	1
0,00	0	0	⋯	1	2
0,65	0	1	⋯	1	1
0,11	0	1	⋯	1	2
0,63	0	0	⋯	1	1

Tabel 6. Data uji

sex	steroid	…	protime	histology	class
0	0	…	0,76	1	2

Berdasarkan tabel 5 di atas terdapat 15 data yang mana 10 data class 1, 5 data class 2. Langkah berikutnya setelah data latih dan data uji telah di ketahui maka mencari jarak terdekat.

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

$D(x,y)$

$$= \sqrt{\begin{matrix}(0{,}22-0{,}89)^2+(1-0)^2+(0-0)^2+(1-1)^2+(1-0)^2+(1-0)^2+(1-1)^2+\\(0-0)^2+(1-1)^2+(1-1)^2+(1-0)^2+(1-1)^2+(1-1)^2+(0{,}14-0{,}08)^2+\\(0{,}11-0{,}06)^2+(0-0{,}01)^2+(0{,}76-0{,}81)^2+(0{,}76-0{,}76)^2+(0-1)^2\end{matrix}}$$

Jarak = 2,34

E. Implementasi ***K****-Nearest Neighbor*

SMOTE adalah teknik membuat data sintetis dengan cara membangkitkan data yang minoritas. Persamaan yang digunakan dalam teknik ini $X_{syn} = X_i + (X_{knn} - X_i)\, x\, \delta$, dimana δ adalah nilai random range 0 sampai 1.

Tabel 7. Data baru yang terbentuk

age	sex	steroid	…	protime	histology	class
0,80	0,2	0	…	0,61	1	2
0,72	0	0	…	0,61	1	2
0,72	0	0,2	…	0,61	1	2

Setelah data terbentuk langkah selanjutnya mencari jarak terdekat *K-Nearest Neighbor* SMOTE.

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

$D(x,y)$

$$= \sqrt{\begin{matrix}(0{,}89-0{,}80)^2+(0-0{,}2)^2+(0-0)^2+(1-0{,}8)^2+(0-0{,}2)^2+\\(0-0{,}2)^2+(1-0{,}8)^2+(0-0)^2+(1-0{,}8)^2+(1-0{,}8)^2+\\(0-0{,}2)^2+(1-0{,}8)^2+(1-0{,}8)^2+(0{,}08-0{,}07)^2+(0{,}06-0{,}05)^2+\\(0{,}01-0{,}01)^2+(0{,}81-0{,}66)^2+(0{,}76-0{,}61)^2+(1-1)^2\end{matrix}}$$

Jarak = 0,68

4. HASIL DAN PEMBAHASAN

A. Hasil Klasifikasi ***K****-Nearest Neighbor* Tanpa SMOTE

Pada proses klasifikasi dilakukan dengan menggunakan Algoritma *K-Nearest Neighbor* tanpa SMOTE. Pengujian dilakukan menggunakan *Kfold* dengan nilai K 2, 4, dan 5. Klasifikasi ini dilakukan untuk mengetahui nilai akurasi, presisi, dan *recall*. Dalam pengujian menggunakan *Kfold* 2 menggunakan algoritma *K-Nearest Neighbor* tanpa SMOTE mendapatkan hasil terbesar dengan akurasi 89,66%, presisi 78,12%, dan *recall* 94,44%.

Tabel 8. *Confusion matrix* algoritma *K-Nearest Neighbor* tanpa SMOTE *Kfold* 2

Kelas prediksi	Kelas aktual	
	Positive (1)	Negative (0)
Positive (1)	6	5
Negative (0)	6	54

Tabel 9. Hasil klasifikasi pada algoritma *K-Nearest Neighbor* tanpa SMOTE

Skenario	*Kfold*	Tahap ke-	Akurasi	Presisi	*Recall*
1	2	1	84,51%	70,76%	72,27%
		2	78,87%	67,37%	64,64%
2	4	1	83,33%	62,50%	91,18%
		2	80,56%	78,12%	64,81%
		3	74,29%	62,50%	61,34%
		4	82,86%	73,21%	73,21%
3	5	1	89,66%	70,00%	94,44%
		2	86,21%	77,56%	67,92%
		3	67,86%	61,36%	58,77%
		4	75,00%	61,30%	59,85%
		5	85,71%	72,73%	81,25%

Skenario yang paling optimal pada dataset penyakit hepatitis terjadi pada skenario ke-3 dengan *Kfold* 5 tahap ke-1 dengan nilai akurasi 89,66 %, presisi 78,12% pada skenario ke-2 dengan *Kfold* 4 pada tahap ke-2, dan *recall* 94,44% pada skenario ke-3 dengan *Kfold* 5 pada tahap ke-1.

B. Hasil Klasifikasi ***K*** *-Nearest Neighbor* SMOTE

Pada proses klasifikasi dilakukan dengan menggunakan Algoritma *K-Nearest Neighbor* SMOTE. Pengujian dilakukan menggunakan *Kfold* dengan nilai K 2, 4, dan 5. Klasifikasi ini dilakukan untuk mengetahui nilai akurasi, presisi, dan *recall*. Pengujian menggunakan *Kfold* 2 dengan algoritma *K-Nearest Neighbor* SMOTE mendapatkan hasil terbesar dengan akurasi 93,48%, presisi 95,00%, dan *recall* 92,11%.

Tabel 10. *Confusion matrix* algoritma *K-Nearest Neighbor* SMOTE *Kfold* 2

Kelas prediksi	Kelas aktual	
	Positive (1)	Negative (0)
Positive (1)	57	2
Negative (0)	12	45

Tabel 11. Hasil klasifikasi pada algoritma *K-Nearest Neighbor* SMOTE

Skenario	*Kfold*	Tahap ke-	Akurasi	Presisi	*Recall*
1	2	1	87,93%	89,18%	87,78%
		2	83,62%	84,07%	83,72%
2	4	1	87,93%	88,47%	88,70%
		2	91,38%	92,17%	90,48%
		3	91,38%	91,83%	91,38%
		4	84,48%	86,40%	84,88%

		1	89,36%	91,07%	89,58%
		2	87,23%	87,32%	87,45%
3	5	3	93,48%	95,00%	92,11%
		4	84,78%	85,62%	85,96%
		5	86,96%	90,00%	86,36%

Skenario yang paling optimal pada dataset penyakit hepatitis terjadi pada skenario ke-3 dengan *Kfold* 5 tahap ke-3 dengan nilai akurasi 93,48 %, presisi 95,00 % pada skenario ke-3 dengan *Kfold* 5 pada tahap ke-3, dan *recall* 92,11 % pada skenario ke-3 dengan *Kfold* 5 pada tahap ke-3.

5. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan diatas maka dapat diambil kesimpulan dari penelitian pengaruh teknik SMOTE terhadap prediksi harapan hidup penderita penyakit hepatitis menggunakan Algoritma *K-Nearest Neighbor* adalah:

A. Klasifikasi *K-Nearest Neighbor* tanpa SMOTE diperoleh hasil terbaik terdapat pada skenario ke-3 dengan *Kfold* 5 tahap ke-1 dengan nilai akurasi 89,66%, presisi 78,12% pada skenario ke-2 dengan *Kfold* 4 pada tahap ke-2, dan *recall* 94,44% pada skenario ke-3 dengan *Kfold* 5 pada tahap ke-1.
B. Klasifikasi *K-Nearest Neighbor* SMOTE diperoleh hasil terbaik terdapat pada skenario ke-3 dengan *Kfold* 5 tahap ke-3 dengan nilai akurasi 93,48%, presisi 95,00% pada skenario ke-3 dengan *Kfold* 5 pada tahap ke-3, dan *recall* 92,11% pada skenario ke-3 dengan *Kfold* 5 pada tahap ke-3.

Beberapa hal yang dapat dikembangkan dalam penelitian ini guna memperoleh hasil yang lebih baik antara lain memperbolehkan peneliti selanjutnya untuk membandingkan dengan metode klasifikasi lain atau menambahkan undersampling untuk perbandingan.

6. DAFTAR PUSTAKA

Ali, A., Shamsuddin, S.M., & Anca L. R. (2015). "Classification with class imbalance problem: A review." *International Journal of Advances in Soft Computing and its Applications 7(3)* : 176–204.

Avila, D., Bussonier, M., & Corlay, S. (2020). Jupyter. Jupyter.org.

Bustami. (2010). "Penerapan Algoritma *Naïve Bayes* untuk Mengklasifikasi Data Nasabah." *TECHSI: Jurnal Penelitian Teknik Informatika* 4: 127–46.

Han. (2006). Classification Clasification.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In Data Mining: Concepts and Techniques. https://doi.org/10.1016/C2009-0-61819-5.

Hidayah, U.N., Oktavianto, H., & Muharom, L.A. (2021). "Analisis Metode K-Nearest Neighbor Terhadap Klasifikasi Data Pasien Penderita Gagal Jantung."

Hokya, S. (2013). "Buku Panduan Pemrograman Python." Buku 84: 487–92. http://ir.obihiro.ac.jp/dspace/handle/10322/3933.

Ismail, A. M. (2018). "Cara Kerja Algoritma K-Nearest Neighbor (K-NN)." Medium.Com (August 2018): Artificial Intelligence. https://medium.com/bee-solution-partners/cara-kerja-algoritma-k-nearest-neighbor-k-nn-389297de543e.

Kemenkes RI. (2015). "Penanggulangan Hepatitis Virus." Peraturan Menteri Kesehatan Republik Indonesia Nomor 53 Tahun 2015 Tentang Penanggulangan Hepatitis Virus Nomor 1126 (879): 1–41.

Khomsah, S. (2018). "Prediksi Harapan Hidup Penderita Hepatitis Kronik Menggunakan Metode-Metode Klasifikasi." *Seminar Nasional Informatika Medis*: 38–45.

Ramdhani, A., Isnanto, R. R., & Windasari, I. P. (2015). "Pengembangan Sistem Pakar Untuk Diagnosis Penyakit Hepatitis Berbasis Web Menggunakan Metode Certainty Factor." *Jurnal Teknologi dan Sistem Komputer* 3(1): 58.

Sari, H. P., Indriastuti, D., Asrul, M., & Elyasari E. (2019). "Perbedaan Pengetahuan Pre dan Post Pendidikan Kesehatan Pada Penghuni Lapas Tentang Risiko Kejadian Viral Hepatitis Di Lapas Perempuan Kelas III." *Jurnal Keperawatan* 2(3): 9–16. https://stikesks-kendari.e-journal.id/JK/article/view/259.

Septiani, W. D. (2017). "Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis." *Jurnal Pilar Nusa Mandiri Volume 13 No.1*: 76–84.

Sulastri, S., Hadiono, K., & Anwar, M. T. (2020). "Analisis Perbandingan Klasifikasi Prediksi Penyakit Hepatitis Dengan Menggunakan Algoritma K-Nearest Neighbor, *Naïve Bayes* dan *Neural Network*." *Dinamik* 24(2): 82–91.

Saxena, K., Khan, Z., & Singh, S. (2014). Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm. *International Journal of Computer Science Trends and Technology* (IJCST), 2(4), 36-43.

Syukron, M., Santoso, R., & Widiharih, T. (2020). "Perbandingan Metode Smote Random Forest Dan Smote Xgboost untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data." *Jurnal Gaussian* 9(3): 227–36.

Ting, K. M. (2017). "Confusion Matrix." Encyclopedia of Machine Learning and Data Mining (October): 260–260.

Wibowo, A. P., & Jumiati, E. (2018). "Sentiment Analysis Masyarakat Pekalongan Terhadap Pembangunan Jalan Tol Pemalang-Batang di Media Sosial." IC-Tech XIII (0285): 42–48. http://ejournal.stmik-wp.ac.id/.

Widiarsana, O., Putra, N.W., & Budiyasa (2011). "Data Mining: Metode Clasification K-Nearest Neighbor (KNN)." Program Studi Teknologi Informasi Universitas Udayana.

Widyatama & Suprapty. (2018). "Bab II Landasan Teori." *Journal of Chemical Information and Modeling* 53(9): 1689–99.