Terakreditasi SINTA Peringkat 3
Surat Keputusan Direktur Jenderal Pendidikan Tinggi, Riset, dan Teknologi Nomor 225/E/KPT/2022
masa berlaku mulai Vol.7 No. 1 tahun 2022 s.d Vol. 11 No. 2 tahun 2026

Terbit online pada laman web jurnal:
http://publishing-widyagama.ac.id/ejournal-v2/index.php/jointecs

JOINTECS

JOINTECS
(Journal of Information Technology and Computer Science)

Vol. 8 No. 1 (2023) 33 - 40

e-ISSN:2541-6448 p-ISSN:2541-3619

Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes

Nabilah Sharfina[1], Nur Ghaniaviyanto Ramadhan[2*]
Program Studi Rekayasa Perangkat Lunak, Fakultas Informatika, Institut Teknologi Telkom Purwokerto
[1]19104025@ittelkom-pwt.ac.id, [2]ghani@ittelkom-pwt.ac.id

Abstract

According to WHO, around 71 million people were infected with the Hepatitis C virus in 2019. However, only 49.7% of people are aware of Hepatitis C. Early prevention is essential to minimize the possibility of something terrible. To maximize the efforts of medical experts in minimizing the risk of transmission, a program was created that is capable of classifying Hepatitis C with an automatic detection system using a machine learning model. Random Forest was chosen because it can handle outlier and imbalance data so that it can produce high accuracy values and can identify important features. Naïve Bayes was chosen because of its simple algorithm, but capable of producing high-accuracy values. After testing both models, the confusion matrix formula calculates the prediction results. The test results show that applying the Random Forest model without SMOTE is 93%, and Naïve Bayes without SMOTE is 88%. Due to the data imbalance in the dataset, an oversampling technique is performed using the SMOTE method. The test results were obtained by applying the Random Forest model with a SMOTE of 98% and Naïve Bayes with a SMOTE of 89%.

Keywords: ***hepatitis c; random forest; naïve bayes; SMOTE; confusion matrix.***

Abstrak

Menurut WHO, orang yang terinfeksi virus Hepatitis C tercatat sekitar 71 juta orang pada 2019. Hanya 49,7% orang yang menyadari adanya penyakit Hepatitis C. Pencegahan dini penting dilakukan untuk meminimalisir kemungkinan buruk terjadi. Untuk memaksimalkan upaya ahli medis dalam meminimalisir risiko penularan, dibuat program yang mampu mengklasifikasikan penyakit Hepatitis C dengan sistem deteksi otomatis menggunakan model *machine learning*. *Random Forest* dipilih karena mampu menangani *outlier* dan *imbalance data* sehingga mampu menghasilkan nilai akurasi yang tinggi serta mampu mengidentifikasi fitur-fitur yang penting. *Naïve Bayes* dipilih karena algoritmanya yang sederhana, namun mampu menghasilkan nilai akurasi tinggi. Setelah dilakukan pengujian pada kedua model, dilakukan perhitungan terhadap hasil prediksi menggunakan formula *confusion matrix*. Hasil pengujian menunjukkan dengan menerapkan model *Random Forest* tanpa SMOTE sebesar 93% dan *Naïve Bayes* tanpa *SMOTE* sebesar 88%. Sehubungan dengan adanya *imbalance data* pada *dataset*, maka dilakukan teknik *oversampling* menggunakan metode SMOTE. Hasil pengujian yang diperoleh dari menerapkan model *Random Forest* dengan SMOTE sebesar 98% dan *Naïve Bayes* dengan SMOTE sebesar 89%.

Kata kunci*:* ***hepatitis c; random forest; naïve bayes; SMOTE; confusion matrix.***

Diterima Redaksi : 19-01-2023 | Selesai Revisi : 07-06-2023 | Diterbitkan Online : 10-06-2023

1. Pendahuluan

Virus Hepatitis merupakan gejala peradangan pada organ hati. Timbulnya penyakit ini disebabkan oleh infeksi yang tidak terdiagnosis selama beberapa dekade, konsumsi senyawa beracun (misalnya obat-obatan, alkohol), atau autoimun. Virus Hepatitis digolongkan menjadi 5 jenis yaitu tipe A, B, C, D, dan E. Menurut pernyataan dari WHO di tahun 2019 bahwa virus Hepatitis B dan C memiliki dampak yang buruk bagi ratusan juta jiwa karena menyebabkan timbulnya gejala awal dari penyakit sirosis dan kanker hati. Sebagian besar virus Hepatitis C ditularkan melalui transfusi darah dan produk darah yang terkontaminasi di luar prosedur medis, bisa juga melalui kontak seksual namun jarang terjadi. Belum ada vaksin untuk penyakit jenis ini [1]. Infeksi virus Hepatitis C (HCV) adalah salah satu penyebab utama munculnya diagnosa penyakit hati kronis di seluruh dunia. Sekitar 71 juta orang yang terdiagnosis kronis di seluruh dunia, namun kesadaran masyarakat masih rendah akan hal tersebut [2]. Menurut hasil survei yang dilakukan oleh lembaga *the Nation's Mobile Health Survey* (NHANES) di Amerika Serikat pada tahun 2001 – 2008, tercatat ada sebanyak 393 orang dari 30.140 orang terjangkit infeksi HCV. Diketahui dari 170 orang yang mampu ditindaklanjuti selanjutnya hanya 49,7% yang mana mereka mengetahui status infeksi HCV. Perkiraan di wilayah Eropa mengatakan bahwa hanya 10% – 40% dari orang-orang dengan diagnosis positif HCV secara individu menyadari adanya status infeksi tersebut. Lembaga *American Association for the Study of Liver Disease* (AASLD) memberikan panduan yang komprehensif dan diperbarui secara berkala faktor risiko, pengujian, evaluasi, pemantauan infeksi HCV, dan perkembangan baru dalam pengobatan [3].

Segala aktivitas manusia di era modern tentu tidak lepas dari penggunaan teknologi seperti komputer, laptop, ponsel pintar, dan lainnya. Seiring berkembangnya teknologi, komputer dilatih menggunakan bahasa pemrograman yang dapat belajar dari kumpulan data untuk melakukan tugas tertentu seperti mengenali pola, membuat prediksi, dan membuat keputusan tanpa melalui program secara eksplisit disebut *machine learning*. Pada penelitian ini, dilakukan klasifikasi terhadap penyakit Hepatitis C menggunakan *machine learning*.

Penelitian tentang identifikasi prediktor sosiodemografi dan klinis perkembangan kaskade perawatan Hepatitis C untuk pasien dalam kelompok kelahiran 1945-1965 di Amerika Serikat bagian selatan menggunakan model *Decision Tree* dan *Random Forest*, yang berjudul "*Predictors of Progression Through the Cascade of Care to a Cure for Hepatitis C Patients using Decision Trees and Random Forests*" oleh Jasmine Ye Nakayama, Joyce Ho, Emily Cartwright, Roy Simpson, dan Vicki Stover Hertzberg pada tahun 2021. Hasil akurasi pada kaskade keterkaitan dengan perawatan menggunakan *Decision Tree* sebesar 64% dan *Random Forest* sebesar 55%, hasil akurasi pada kaskade inisiasi pengobatan antivirus menggunakan *Decision Tree* sebesar 76% dan *Random Forest* sebesar 73%, serta hasil akurasi pada kaskade penyembuhan virologi menggunakan *Decision Tree* sebesar 75% dan *Random Forest* sebesar 55% [4]. Adapun penelitian lain yang membahas tentang pengembangan model *Decision Tree* dan *Naïve Bayes* dalam memprediksi pasien Hepatitis C, dengan judul penelitian "Analisis Penerapan Metode *Ensembled Learning Decision Tree* pada Klasifikasi Virus Hepatitis C" oleh Rifqi Alfinnur Charisma, Sofiyudin Pamungkas, Rifqi Akmal Saputra, Nur Ghaniaviyanto Ramadhan, dan Faisal Dharma Adhinata pada tahun 2022. Berhasil memperoleh nilai akurasi tinggi pada pengujian model *Decision Tree* sebesar 93%, sedangkan *Naïve Bayes* sebesar 90% [5]. Penelitian selanjutnya yang berjudul "KLASIFIKASI HEPATITIS C VIRUS MENGGUNAKAN ALGORITMA C4.5" oleh Susanto dan Nuri pada tahun 2022, membahas tentang klasifikasi penyakit Hepatitis C menggunakan model C4.5 biasa dan model C4.5 berbasis *Adaboost*. Hasil akurasi dari penggunaan model C4.5 biasa sebesar 95,71%, sedangkan pada model C4.5 berbasis *Adaboost* sebesar 96,55% [6].

Untuk mengetahui performa klasifikasi dari hasil uji model, bisa menggunakan tabel perhitungan *Confusion Matrix*. *Confusion Matrix* terdiri dari 4 tipe karakteristik dalam mengukur nilai aktual dan nilai prediksi suatu klasifikasi di antaranya *true positive, true negative, false positive*, dan *false negative*. Secara konsep perhitungan terdiri dari *accuracy, precision*, dan *recall* [7]. Selama dilakukan pemrosesan data, ditemukan adanya kondisi jumlah target atribut kolom terlihat tidak seimbang antara satu dengan yang lain. Kondisi seperti itu, wajar terjadi pada sebuah *dataset*. Apabila hal tersebut dibiarkan, kemungkinan cenderung terjadi risiko salah mengklasifikasi data ketika dilakukan pengujian terhadap model yang diusulkan. Solusi untuk menghindari risiko tersebut, dilakukan percobaan ulang dengan membuat data buatan menggunakan teknik *Synthetic Minority Oversampling Technique* (SMOTE).

Menurut penelitian yang dilakukan oleh Laila Qadrini dan Hikmah Megasari pada tahun 2022 yang berjudul "*Oversampling, Undersampling,* SMOTE SVM dan *Random Forest* pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017", dilakukan klasifikasi pada studi kasus siswa SMA atau sederajat yang berhak menerima bidikmisi di Jawa Timur pada tahun 2017. Model yang diusulkan adalah SVM dan *Random Forest*. Terdapat kondisi data yang digunakan tidak seimbang jumlahnya, sehingga dilakukan teknik *undersampling, oversampling*, dan SMOTE. Hasil akurasi yang didapat dari penerapan model SVM setelah *oversampling* sebesar 0,591, setelah *undersampling* sebesar 0,032, dan setelah SMOTE sebesar 0,761. Hasil akurasi yang didapat dari penerapan model *Random Forest* setelah *oversampling* sebesar 0,735, setelah *undersampling*

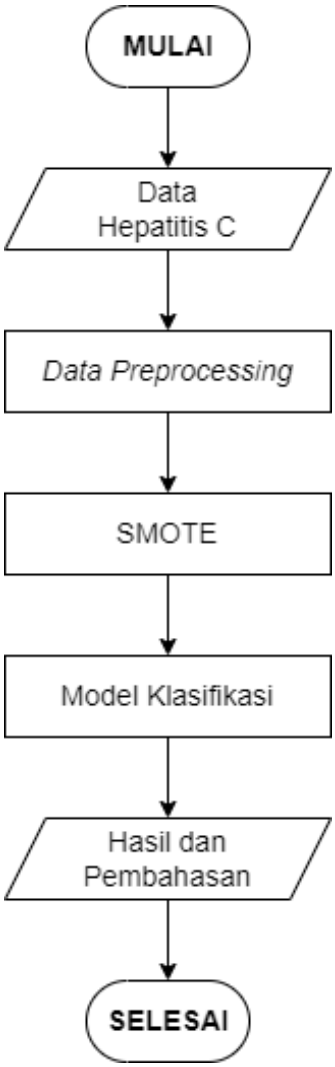
sebesar 0,199, dan setelah SMOTE sebesar 0,78 [8]. Penelitian berikutnya yang dilakukan oleh Desti Mualfah, Wahyu Fadila, dan Rahmad Firdaus pada tahun 2022 yang berjudul "Teknik SMOTE Untuk Mengatasi I*mbalance Data* pada Deteksi Penyakit *Stroke* Menggunakan Algoritma *Random Forest*", dilakukan deteksi stroke dengan menerapkan model *Random Forest* dan teknik SMOTE untuk menangani data tidak seimbang. Performa model *Random Forest* sebelum SMOTE menghasilkan akurasi sebesar 98%, presisi sebesar 69%, *recall* sebesar 51%, dan *f1-score* sebesar 51%. Performa model *Random Forest* sesudah SMOTE menghasilkan akurasi sebesar 91%, presisi sebesar 92%, *recall* sebesar 91%, dan *f1-score* sebesar 91%. Terjadi peningkatan terhadap presisi, *recall*, dan *f1*-score [9].

Berdasarkan uraian latar belakang masalah dan penelitian terdahulu, maka peneliti akan mengklasifikasi virus Hepatitis C menggunakan model *Random Forest* dan *Naïve Bayes*. Penerapan *Random Forest* dianggap memiliki keunggulan pada sisi penanganan *outlier*, efektif dalam menangani *imbalance data*, waktu dalam memproses komputasi lebih cepat, dan hasil akurasi yang tinggi [10]. Selain itu, penerapan metode lain seperti *Naïve Bayes* dianggap memiliki kelebihan yaitu perhitungan algoritmanya yang lebih sederhana namun mampu menghasilkan nilai akurasi yang tinggi untuk klasifikasi data [11]. Penerapan teknik SMOTE pada penelitian ini bertujuan untuk mengontrol distribusi data yaitu mengubah jumlah data dari kelas minoritas menjadi sebanyak jumlah data kelas mayoritas [12]. Pada penerapannya, SMOTE tidak mengurangi jumlah data sehingga informasi penting tidak hilang [13]. Sumber penelitian yang menjadi acuan utama adalah penelitian yang dilakukan oleh tim Rifqi Alfinnur Charisma, karena menggunakan model penelitian yang berbeda pada *dataset* yang sama. Tujuannya untuk memanfaatkan kemajuan teknologi seperti *machine learning* di bidang ilmu kesehatan. Diagnosis dan pengobatan dini penting dilakukan guna mencegah dan meningkatkan kemungkinan penyembuhan pasien Hepatitis C. Harapannya mampu menghasilkan nilai akurasi yang tinggi saat pengujian model sehingga membantu tim ahli medis ketika melakukan identifikasi pasien apabila memiliki tingkat risiko tinggi, agar segera ditindakan lanjut guna mengurangi kemungkinan berkembangnya komplikasi tersebut.

2. Metode Penelitian

Jelaskan metode preparasi dan teknik karakterisasi yang digunakan. Jelaskan dengan ringkas, tetapi tetap akurat seperti ukuran, volume, replikasi dan teknik pengerjaan. Untuk metode baru harus dijelaskan secara rinci agar peneliti lain dapat mereproduksi percobaan

Adapun uraian terkait alur metodologi penelitian yang dilakukan demi mencapai tujuan analisis penelitian, bisa dilihat pada Gambar 1. Proses dimulai dengan melakukan pra-pemrosesan data dengan menangani imbalanced data menggunakan SMOTE, lalu dilakukan klasifikasi menggunakan model *random forest* dan *naïve bayes* untuk dilakukan perbandingan.



Gambar 1. Metodologi Penelitian Virus Hepatitis C

2.1. Data Hepatits C

Pada tahap ini, data yang digunakan sebagai bahan penelitian didapatkan melalui studi literatur penelitian terdahulu. Diperoleh data yang berkaitan dengan topik yang diangkat pada penelitian, sebelum dilakukan klasifikasi menggunakan program *machine learning*. Data pasien diagnosis Hepatitis C berasal dari situs web *Kaggle*, berupa format dokumen (*.csv*) yang terdiri dari 615 baris data dengan 13 atribut kolom [5], bisa dilihat pada Gambar 2 sebagai berikut.

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
2	0=Blood Donor	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
4	0=Blood Donor	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7

Gambar 2. Data Darah Pasien Diagnosis Hepatitis C

Penjelasan dari masing-masing nama atribut kolom tertulis pada Tabel 1, sebagai berikut.

Tabel 1. Deskripsi Informasi Data

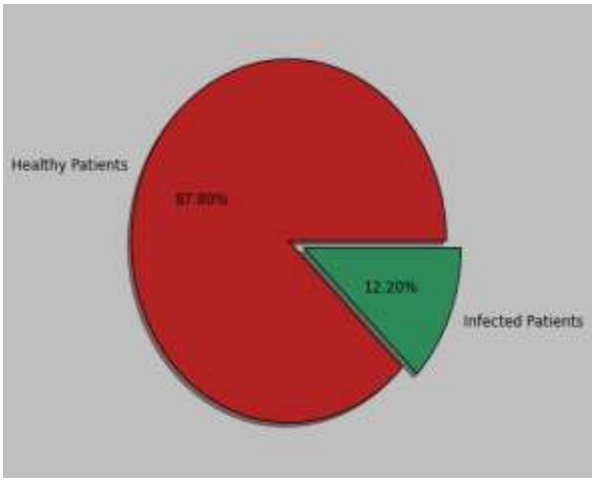
Nama Atribut Kolom	Deskripsi Data	Kategori Tipe Data
Category	Target fitur uji	Objek
Age	Usia pasien	Numerik (*integer*)
Sex	Jenis kelamin pasien	Objek
ALB	Jumlah *albumin*	Numerik (*float*)
ALP	Jumlah *alkaline phosphatase*	Numerik (*float*)
ALT	Jumlah *alanine transaminase*	Numerik (*float*)
AST	Jumlah *aspartate aminotransferase*	Numerik (*float*)
BIL	Jumlah *bilirubin*	Numerik (*float*)
CHE	Jumlah *cholinesterase*	Numerik (*float*)
CHOL	Jumlah *cholesterol*	Numerik (*float*)
CREA	Jumlah *creatine*	Numerik (*float*)
GGT	Jumlah *gamma-glutamyl transferase*	Numerik (*float*)
PROT	Jumlah *protein*	Numerik (*float*)

2.2. *Data Preprocessing*

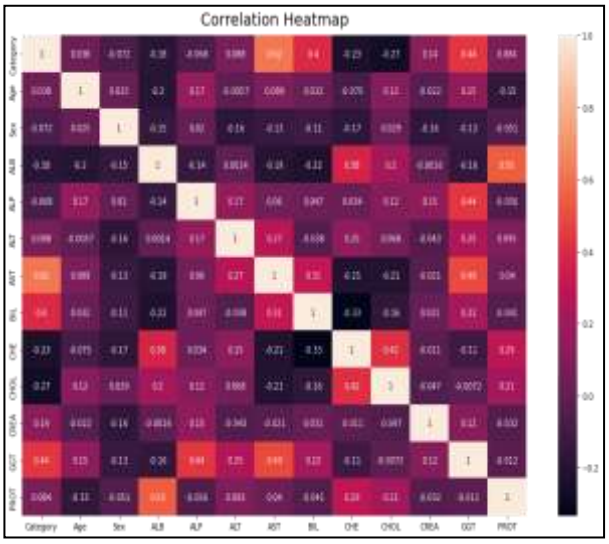
Pada tahap ini, dilakukan optimasi pada program sebelum melangkah ke proses selanjutnya. Sehubungan pada pengolahan data ditemukan adanya nilai yang kosong, maka pada atribut kolom "*ALB*", "*ALP*", "*ALT*", "*CHOL*" dan "*PROT*", maka dapat diatasi dengan cara mengisi nilai rata-rata agar tidak terjadi bias [14][15]. Selanjutnya, dilakukan pengecekan terhadap duplikasi data, namun tidak ditemukan adanya duplikasi data. Setelah itu, mengubah tipe data objek menjadi numerik (*integer*) pada atribut "*category*" dan "*sex*". Ketika dilakukan analisis terhadap atribut "*category"*, mendapati bahwa atribut tersebut terdiri dari 4 kelas meliputi *blood donor, suspect blood donor, hepatitis, fibrosis*, dan *cirrhosis*. Untuk atribut "*category*" dikelompokkan menjadi 2 kategori kelas yaitu "0=*Blood Donor*" = 0, "0s=*suspect Blood Donor*" = 0, "1=*Hepatitis*" = 1, "2=*Fibrosis*" = 1, dan "3=*Cirrhosis*" = 1. Peneliti mendefinisikan kelas 0 menjadi kategori kelas pasien dengan diagnosis sehat (*healthy patients*) ada sebanyak 540 data, sedangkan kelas 1 menjadi kategori kelas pasien dengan diagnosis terpapar infeksi (*infected patients*) ada sebanyak 75 data. Secara persentase bisa dilihat visualisasi diagram *pie chart* pada Gambar 3.

Sedangkan, untuk atribut "*sex*" dikelompokkan menjadi 2 kategori kelas pula yakni "m" = 1 dan "f" = 2. Apabila semua atribut sudah berubah menjadi tipe data numerik, maka tahap selanjutnya adalah mengecek korelasi antar atribut menggunakan visualisasi *heatmap*, seperti pada Gambar 4.

Sebelum dilakukan pemisahan data, dilakukan proses *encoding* menggunakan *LabelEncoder*, digunakan untuk fitur kategorik dengan tipe data ordinal seperti kolom "*category*". Selanjutnya, dalam memudahkan proses *training*, maka dibuat variabel X yang berisi semua fitur pada *dataset* untuk melatih model (*independent variable*), sedangkan variabel Y berisi target variabel (*dependent variable*) [16]. Setelah itu, dilakukan normalisasi data menggunakan *StandardScaler*. Jumlah data sebanyak 615 pasien akan dibagi menjadi 2 dengan proporsi 80% (*data train*) : 20% (*data test*). Hasil dari proses pemisahan data, mendapati sebanyak 492 *data train* dan 123 *data test*.



Gambar 3. Sebaran Kategori Pasien



Gambar 4. Korelasi Data Menggunakan *Heatmap*

2.3. *Synthetic Minority Oversampling Technique* (SMOTE)

Berdasarkan studi literatur, *Synthetic Minority Oversampling Technique* (SMOTE) merupakan algoritma yang bekerja dengan tujuan mengontrol pemerataan distribusi data pada suatu kelas dalam dataset dengan membuat sampel buatan dari kelas minoritas. Konsepnya yaitu memilih sampel secara acak dari kelas minoritas, kemudian menghiung nilai k terdekat sebagai sampel, berikutnya sampel buatan ditambahkan ke kumpulan data yang asli [17][18].

Teknik ini sering digunakan untuk menangani imbalance data, guna membantu perfoma model usulan.

Tabel 2. Sebaran Data SMOTE

Kelas	Data Asli	Data Sampel
0	51	441
1	441	441

Terlihat pada Tabel 2, menunjukkan bahwa penerapan teknik SMOTE berhasil menyeimbangkan jumlah data. Jumlah data yang dilakukan SMOTE adalah pada kelas minor yaitu kelas tidak atau 0. Hal tersebut nantinya saat dilakukan prediksi menggunakan dua model akan memiliki dampak terhadap hasil akurasinya.

2.4. Model Klasifikasi

Pada tahap ini, dilakukan pengujian terhadap 2 model usulan untuk menentukan model terbaik. Hasil pengujian dilihat berdasarkan nilai akurasi, presisi, dan recall. Model terbaik yang dihasilkan diharapkan dapat diterapkan pada kasus prediksi penyakit lainnya.

2.4.1. *Random Forest*

Berdasarkan studi literatur, *Random Forest* adalah jenis algoritma *ensemble learning* yang menggunakan banyak pohon keputusan untuk membuat prediksi dengan mengurangi varian dari pohon keputusan yang baru. Cara kerja dari algoritma pohon keputusan dilakukan dengan memilih kumpulan dari variabel secara acak (fitur). Kumpulan pohon acak tersebut disebut *Random Forest*. *Random Forest* dianggap sebagai satu algoritma klasifikasi yang paling akurat, karena mampu menghasilkan nilai akurasi yang tinggi. Karakteristik lain dari *Random Forest* adalah secara signifikan mampu menangani *imbalance data* dibandingkan dengan model lain serta mampu mengidentifikasi fitur-fitur penting dalam data [19].

2.4.2. *Naïve Bayes*

Berdasarkan studi literatur, *Naïve Bayes* adalah salah satu model yang memiliki perhitungan probabilitas sederhana dan akurat dalam kebutuhan klasifikasi sehingga dapat diterapkan dalam kegiatan penambangan data. *Naïve Bayes* berasal dari teorema *Bayes* yang sering dibahas dalam teori probabilitas. Teorema tersebut menyatakan dua kejadian X dan Y [19], seperti rumus 1.

$$P(X|Y) = \frac{P(Y|X)\ x\ P(X)}{P(Y)} \quad (1)$$

di mana P(X|Y) adalah probabilitas bersyarat dari peristiwa X yang terjadi, jika diketahui bahwa peristiwa Y telah terjadi. P(Y|X) adalah probabilitas bersyarat dari peristiwa Y yang terjadi, jika diketahui bahwa peristiwa X telah terjadi. P(X) adalah yang sebelumnya probabilitas peristiwa X terjadi. P(Y) adalah probabilitas sebelumnya terjadinya peristiwa Y.

3. Hasil dan Pembahasan

Setelah didapat hasil pengujian model, dilakukan perhitungan terhadap tingkat keberhasilan klasifikasi program. Demi mencapai tujuan tersebut, maka digunakan formula *confusion matrix* [20]. Perhitungan menggunakan *confusion matrix* digunakan untuk menghitung nilai akurasi, presisi, dan recall.

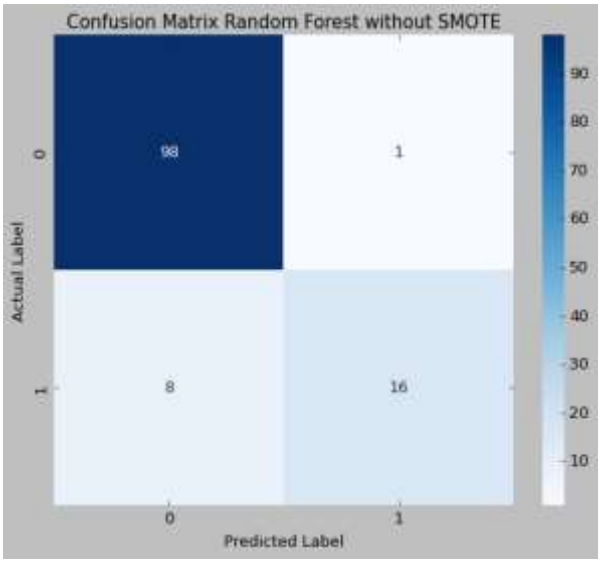
Tabel 3. *Confusion Matrix*

		Predicted Label	
		0 (Sehat)	1 (Terinfeksi)
Actual Label	0 (Sehat)	*True Positive* (TP)	*False Positive* (FP)
	1 (Terinfeksi)	*False Negative* (FN)	*True Negative* (TN)

Pada Tabel 3 merupakan formula untuk menghitung hasil prediksi. Hasil analisa penerapan dua model *naïve bayes* dan *random forest* dengan menggunakan dan tanpa menggunakan SMOTE juga akan dihitung pada bab ini. Perhitungan menggunakan *confusion matrix* dapat mengetahui jumlah prediksi yang salah dan benar.

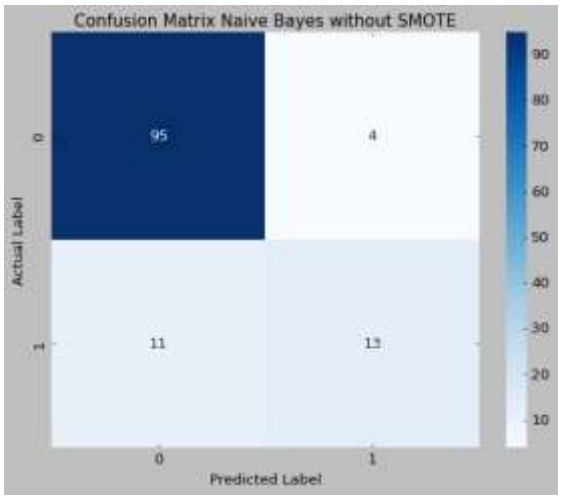
3.1. Implementasi Model tanpa SMOTE

Terlihat pada Gambar 5, menampilkan hasil prediksi klasifikasi dari implementasi model *Random Forest* tanpa teknik SMOTE. Didapatkan masih memiliki kesalahan dalam melakukan klasifikasi dengan total nilai kesalahan 9. Sehingga, perlu dilakukan penerapan SMOTE untuk mengurangi kesalahan klasifikasi tersebut.



Gambar 5. Tabel *Confusion Matrix Random Forest* tanpa SMOTE

Terlihat pada Gambar 6, menampilkan hasil prediksi klasifikasi dari implementasi model *Naïve Bayes* tanpa teknik SMOTE. Naïve bayes memiliki kesalahan dalam melakukan prediksi lebih banyak daripada *random forest*. Hal ini memungkinkan model *naïve bayes* belum terlalu baik dalam prediksi penyakit hepatitis c.



Gambar 6. Tabel *Confusion Matrix Naive Bayes* tanpa SMOTE

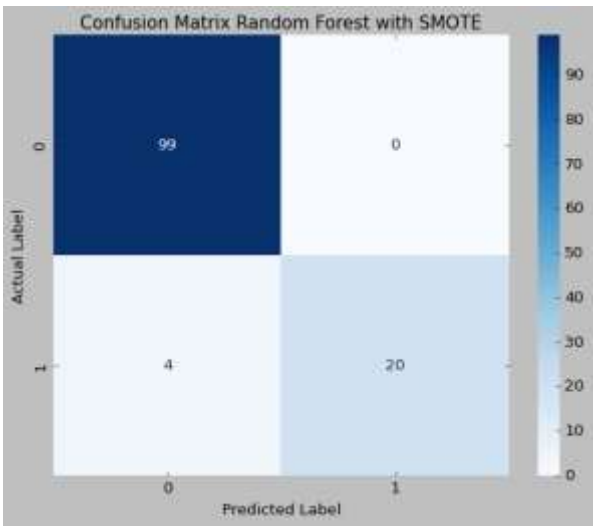
Tabel 4. Hasil Uji Model tanpa SMOTE

Model	*Accuracy*	*Precision*	*Recall*
Random Forest	93%	93%	83%
Naïve Bayes	88%	83%	75%

Pada Tabel 4, merupakan hasil uji perhitungan prediksi kedua model tanpa teknik SMOTE. Hasil confusion matrix sudah terlihat tinggi, namun sebenarnya masih dapat ditingkatkan lagi dengan menerapkan SMOTE. Hasil penerapan SMOTE dapat dilihat pada Tabel 5.

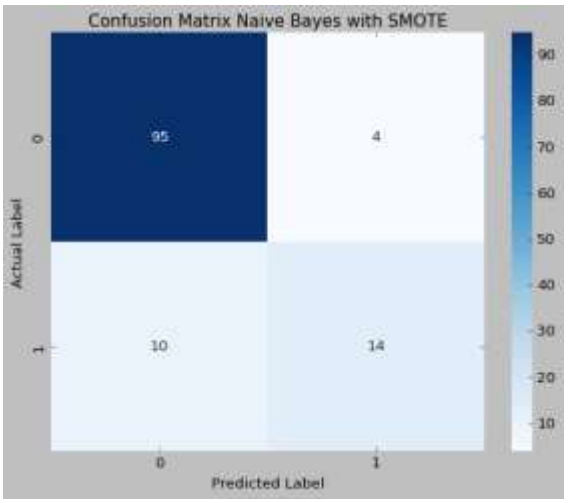
3.2. Implementasi Model Menggunakan SMOTE

Pada program terdapat *imbalance data*, terlihat dengan bantuan visualisasi diagram batang. Untuk mengatasi hal tersebut, maka dilakukan proses *oversampling* menggunakan SMOTE agar jumlah data seimbang satu sama lain. Terlihat pada Gambar 7, menampilkan hasil prediksi klasifikasi dari implementasi model *Random Forest* menggunakan SMOTE.



Gambar 7. Tabel *Confusion Matrix Random Forest* dengan SMOTE

Terlihat pada Gambar 8, menampilkan hasil prediksi klasifikasi dari implementasi model *Naïve Bayes* menggunakan SMOTE. Model naïve bayes masih memiliki kesalahan dalam melakukan klasifikasi sebanyak 4 data dan 10 data. Sehingga, total kesalahan data yang dilakukan klasifikasi oleh model naïve bayes sebesar 14 data. Hal tersebut yang menyebabkan model naïve bayes akurasi masih rendah.



Gambar 8. Tabel *Confusion Matrix Naive Bayes* dengan SMOTE

Tabel 5. Hasil Uji Model dengan SMOTE

Model	*Accuracy*	*Precision*	*Recall*
Random Forest	97%	98%	92%
Naïve Bayes	89%	84%	77%

Pada Tabel 5, merupakan hasil uji perhitungan prediksi kedua model menggunakan SMOTE. Hasil menunjukkan dengan menerapkan SMOTE bahwa model random forest lebih unggul dengan selisih 8%. Hal tersebut menunjukkan bahwa model random forest powerful dalam melakukan klasifikasi penyakit hepatitis C.

Tabel 6. Perbandingan Hasil Penelitian Sebelumnya

Model	*Accuracy*	*Precision*	*Recall*
Random Forest	**97%**	**98%**	**92%**
Naïve Bayes	89%	84%	77%
[4]	55%	-	-
[5]	90%	-	-
[6]	96.5%		
[8]	78%		

Sedangkan, pada Tabel 6 merupakan perbandingan dengan penelitian sebelumnya. Pada penelitian-penelitian sebelumnya hasil akurasi dapat ditingkatkan dengan selisih terbanyak 42% [4]. Hal ini menunjukkan bahwa penerapan SMOTE pada data yang *imbalanced* memiliki dampak terhadap hasil akurasi klasifikasi.

4. Kesimpulan

Berdasarkan hasil dan pembahasan, klasifikasi penyakit Hepatitis C menggunakan model *Random Forest* dan *Naïve Bayes*, dapat disimpulkan bahwa penggunaan

model *Random Forest* tanpa SMOTE mampu menghasilkan nilai akurasi sebesar 93%, sedangkan penggunaan model *Random Forest* menggunakan SMOTE mampu meningkatkan nilai akurasi sebesar 98%. Penggunaan SMOTE juga mampu mengurangi risiko *error* terhadap prediksi data, terlihat pada hasil *confusion matrix*. Metode yang diusulkan, berhasil dalam melakukan klasifikasi penyakit Hepatitis C. Pada penelitian dilakukan pula perbandingan dengan model lain yakni *Naïve Bayes*, namun angka persentase akurasi yang dihasilkan lebih rendah dari hasil model *Random Forest.*

Ucapan Terimakasih

Ucapan terima kasih ditujukan kepada Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) dan dosen pembimbing Tugas Akhir, sekaligus rekan peneliti, yang mana telah memberikan dukungan penuh dalam penyusunan jurnal penelitian ini.

Daftar Pustaka

[1] Schwarz, Tanja, et al. "Interventions to increase linkage to care and adherence to treatment for hepatitis C among people who inject drugs: A systematic review and practical considerations from an expert panel consultation." *International Journal of Drug Policy*, vol. 102, no. 4, pp. 103588, 2022, doi: 10.1016/j.drugpo.2022.103588.

[2] J. Pawlotsky *et al.*, "EASL Recommendations on Treatment of Hepatitis C 2018," *J. Hepatol.*, vol. 4, no. 9, 2018, doi: 10.1016/j.jhep.2018.03.026.

[3] A. A. Rabaan *et al.*, "Overview of hepatitis C infection, molecular biology, and new treatment," *J. Infect. Public Health*, vol. 13, no. 5, pp. 773–783, 2020, doi: 10.1016/j.jiph.2019.11.015.

[4] J. Ye, J. Ho, E. Cartwright, R. Simpson, and V. Stover, "Predictors of progression through the cascade of care to a cure for hepatitis C patients using decision trees and random forests," *Comput. Biol. Med.*, vol. 134, no. March, p. 104461, 2021, doi: 10.1016/j.compbiomed.2021.104461.

[5] R. A. Charisma, S. Pamungkas, R. A. Saputra, and N. G. Ramadhan, "Analisis Penerapan Metode Ensembled Learning Decision Tree Pada Klasifikasi Virus Hepatitis C," *J. Comput. Syst. Informatics*, vol. 3, no. 4, pp. 405–409, 2022, doi: 10.47065/josyc.v3i4.2064.

[6] S. Tinggi, T. Pati, P. Korespondensi, and H. C. Virus, "Klasifikasi Hepatitis C Virus Menggunakan Algoritma C4.5" *J. DISPROTEK*, vol. 13, no. 2, pp. 131–136, 2022, doi: 10.34001/jdpt.v12i2.

[7] A. Muslih, M. F. Ahadi, and M. I. Rasyid, "Klasifikasi Kematangan Pada Buah Mangga Garifta Merah dengan Transformasi Ruang Warna HSI," *J. Appl. Informatics Comput.*, vol. 5, no. 2, pp. 117–121, 2021.

[8] L. Qadrini, H. Hikmah, and M. Megasari, "Oversampling, Undersampling, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017," *J. Comput. Syst. Informatics*, vol. 3, no. 4, pp. 386–391, 2022, doi: 10.47065/josyc.v3i4.2154.

[9] D. Mualfah, W. Fadila, and R. Firdaus, "Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 2, pp. 107–113, 2022, doi: 10.37859/coscitech.v3i2.3912.

[10] J. Gaussian, "Perbandingan Metode SMOTE Random Forest dan SMOTE XGBoost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data," *J. Gaussian*, vol. 9, pp. 227–236, 2020.

[11] E. Pusporani, S. Qomariyah, and I. Irhamah, "Klasifikasi Pasien Penderita Penyakit Liver dengan Pendekatan Machine Learning," *Inferensi*, vol. 2, no. 1, p. 25, 2019, doi: 10.12962/j27213862.v2i1.6810.

[12] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.

[13] J. Estublier *et al.*, "Impact of software engineering research on the practice of software configuration management," *ACM Trans. Softw. Eng. Methodol.*, vol. 14, no. 4, pp. 383–430, 2005, doi: 10.1145/1101815.1101817.

[14] A. Alhamad, A. I. S. Azis, B. Santoso, and S. Taliki, "Prediksi Penyakit Jantung Menggunakan Metode-Metode Machine Learning Berbasis Ensemble – Weighted Vote," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, p. 352, 2019, doi: 10.26418/jp.v5i3.37188.

[15] F. Tang and H. Ishwaran, "Random forest missing data algorithms," *Stat. Anal. Data Min. Asa Data Sci. J.*, no. April, pp. 1–15, 2017, doi: 10.1002/sam.11348.

[16] R. R. Santoso, R. Megasari, and Y. A. Hambali, "Implementasi Metode Machine Learning Menggunakan Algoritma Evolving Artificial Neural Network Pada Kasus Prediksi Diagnosis Diabetes," *JATIKOM (Jurnal Apl. dan Teor. Ilmu Komputer)*, vol. 3, no. 2, pp. 85–97, 2020.

[17] J. Homepage, N. Suryana, and R. Tri Prasetio, "Penanganan Ketidakseimbangan Data pada Prediksi Customer Churn Menggunakan Kombinasi SMOTE dan Boosting," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 6, no. 1, pp. 31–37, 2021, [Online]. Available: https://creativecommons.org/licenses/by-sa/4.0/.

[18] N. G. Ramadhan. "Comparative Analysis Of Adasyn-Svm And Smote-Svm Methods On The Detection Of Type 2 Diabetes Mellitus." *Sci. Jour.*

Of. Inform. vol. 8, no. 2, pp. 276-282, 2021. doi: 10.15294/sji.v8i2.32484.

[19] K. Lemons, "A Comparison Between Naïve Bayes and Random Forest to Predict Breast Cancer," *IJURCA Int. J. Undergrad. Res. Creat. Act.*, vol. 12, 2020.

[20] N. G. Ramadhan and F. D. Adhinata, "Teknik Smote Dan Gini Score Dalam Klasifikasi Kanker Payudara," *RADIAL J. Perad. Sains, Rekayasa dan Teknol.*, vol. 9, no. 2, pp. 125–134, 2021, doi: 10.37971/radial.v9i2.229.