ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 15, Nomor 1, Tahun 2025, Halaman 121 - 130
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENERAPAN KLASIFIKASI REGRESI LOGISTIK BINER DAN *ADAPTIVE BOOSTING* MENGGUNAKAN *CLASSIFICATION AND REGRESSION TREES* PADA PREDIKSI PENYAKIT HEPATITIS C

Ellina Dhiya Ulhaq Oktaviani[1*], Rukun Santoso[2*], Tatik Widiharih[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail**: rukunsantoso25@gmail.com

DOI: 10.14710/j.gauss.15.1.121-130

Article Info:
Received: 2024-12-06
Accepted: 2025-12-30
Available Online: 2026-06-21

Keywords:
Hepatitis C; Binary Logistic Regression; Adaptive Boosting; Synthetic Minority Oversampling Technique

Abstract: Chronic liver disease is primarily attributed to the hepatitis C virus. Disorders of liver function can inhibit metabolism and threaten health. Hepatitis C disease must be detected earlier to reduce the risk of spreading it. Data processing using the Binary Logistic Regression and Adaptive Boosting classification methods to predict the category of patients with positive or negative hepatitis C status. Problems with unbalanced data are found in the classification process. Data imbalance can be overcome with the Synthetic Minority Over-Sampling Technique (SMOTE). Data retrieval was obtained from the 2020 UCI (University of California Irvine) Machine Learning Repository regarding data on predictions of hepatitis C patients which were downloaded on October 25, 2022. The results for the accuracy of the classification show that the Binary Logistic Regression method produces an accuracy value of 97,44%, the value sensitivity of 100%, and specificity of 97,17%. The accuracy of the classification produced by the Adaptive Boosting method with an accuracy value of 92,31%, a sensitivity value of 63,64%, and specificity of 100%. Binary Logistic Regression is the best method that can classify hepatitis C status of patients with the highest sensitivity of 100%.

1. PENDAHULUAN

Gangguan pada hati dapat menyebabkan penyakit yang mengancam kesehatan. Salah satu penyebab utama penyakit hati kronis adalah virus hepatitis C. WHO (*World Health Organization*) menyatakan lebih dari 3000 orang meninggal setiap hari karena penyakit hati yang disebabkan oleh virus hepatitis (*World Health Organization*, 2022). Pengurangan risiko penyebaran penyakit hepatitis C dapat dilakukan dengan deteksi dini. Hasil pemeriksaan labolatorium pada pasien dapat digunakan untuk mengklasifikasi karakteristik pasien dengan status positif hepatitis C atau negatif hepatitis C.

Klasifikasi adalah metode statistika berupa pengelompokkan data berdasarkan karakteristiknya dengan terstruktur ke dalam kategori yang ditentukan (Prasetyo, 2012). Permasalahan yang sering ditemukan pada saat klasifikasi yaitu ketidakseimbangan data. Data imbalance dapat ditangani dengan *sampling* data asli, salah satunya menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Metode klasifikasi pada penelitian ini adalah Regresi Logistik Biner dan *Adaptive Boosting*.

Regresi Logistik Biner adalah regresi logistik yang memiliki variabel respon biner. Penelitian terkait Regresi Logistik Biner oleh Suwardika (2017) membahas mengenai prediksi pasien penderita penyakit hepatitis dengan respon meninggal atau hidup tanpa penanganan data *imbalance*. Penelitian tersebut menghasilkan angka akurasi 79,4%. *Adaptive Boosting* diterapkan pada penelitian ini dengan memberikan suatu bobot lebih pada

observasi yang hasil klasifikasinya dengan tepat. Rabbani *et al*. (2021) menggunakan algoritma *Adaptive Boosting* pada klasifikasi data deteksi jatuh. Penelitian tersebut deteksi jatuh dan tidak jatuh telah diklasifikasikan dan menghasilkan nilai akurasi tertinggi di antara 4 rasio yaitu 97,5% pada rasio 20%TR:80%TS, 98,7% pada rasio 30%TR:70%TS, 99,3% pada rasio 40%TR:60%TS dan 100% pada rasio 50%TR:50%TS.

Penanganan data *imbalance* pada penelitian ini menggunakan metode SMOTE supaya algoritma klasifikasi tidak cenderung berfokus pada kelas mayoritas. Penelitian ini bertujuan membandingkan hasil klasifikasi Regresi Logistik Biner dan *Adaptive Boosting* dalam memprediksi status hepatitis C pasien menggunakan nilai sensitivitasnya.

2. TINJAUAN PUSTAKA

Hepatitis C termasuk dalam virus RNA yang menyebabkan hepatitis non-A dan non-B. Virus ini diidentifikasi pertama kali pada April 1989 (Ikatan Dokter Anak Indonesia, 2009). Pasien dapat diketahui terinfeksi hepatitis C apabila terdapat perkembangan antibodi anti-HCV pada tubuhnya (Perhimpunan Peneliti Hati Indonesia, 2017). Upaya untuk mengurangi penularan hepatitis C dapat dilakukan dengan pemeriksaan laboratorium terkait uji fungsi hati dan aktivitas enzim di dalamnya sehingga dapat memprediksi status hepatitis C pasien.

Data *imbalance* menjadi masalah yang sering ditemukan dalam proses klasifikasi, yaitu suatu keadaan terdapat perbedaan jumlah data antara kelas satu dengan kelas lainnya yang tidak seimbang (Ali, *et al*., 2015). Hal ini akan menjadikan data dengan kelas mayoritas memiliki akurasi yang baik, sedangkan data dengan kelas minoritas akan memiliki akurasi yang kurang baik. (Qiong*, et al*., 2016). Pasien dengan hasil laboratorium positif hepatitis C pada penelitian ini memiliki jumlah selisih yang jauh lebih banyak dibanding kelas pasien yang negatif. Teknik penanganan data *imbalance* dalam penelitian ini yaitu *Synthetic Minority Over-Sampling Technique* (SMOTE). Metode SMOTE menghasilkan data buatan yang banyaknya sesuai dengan persentase duplikasi yang ditentukan antara data minor dengan menemukan ketetanggan terdekat sebanyak k secara acak dari setiap data pada kelas minoritas (Chawla, *et al*., 2002).

Pembangkitan data sintetis dilakukan dengan menggunakan persamaan (1).

$$x_{syn_h} = x_i + (x_{knn} - x_i) \times \gamma_h \qquad (1)$$

Penentuan tetangga terdekat dilakukan menggunakan jarak *Euclidian* untuk menghitung perbedaan jarak pada data numerik. Jika terdapat dataset dengan jumlah variabel sebanyak l, maka untuk menghitung jarak antara $\boldsymbol{x^T} = [x_1, x_2, \dots, x_l]$ dan $\boldsymbol{y^T} = [y_1, y_2, \dots, y_l]$ dapat menggunakan persamaan (2).

$$d(\boldsymbol{x_i}, \boldsymbol{y_i}) = \sqrt{\sum_{l=1}^{L} \left(diff(x_{il}, y_{il})\right)^2} \qquad (2)$$

Normalisasi adalah proses memberi skala nilai atribut sehingga bisa berada pada rentang nilai yang ditentukan. Rumus dari metode *Min-Max* dapat dilihat pada persamaan (3).

$$x' = \frac{x - xmin}{xmax - xmin} \qquad (3)$$

Metode Regresi Logistik Biner sering digunakan dalam menghubungkan variabel dependen yang sifatnya biner dengan variabel independen (Hosmer dan Lemeshow, 2000). Model statistik yang dihasilkan dengan variabel respon biner untuk setiap subjek yaitu "sukses" atau "gagal" (Agresti, 2007). Pembentukkan model regresi logistiknya seperti pada persamaan (4).

$$\pi(\boldsymbol{x_i}) = \frac{exp\,(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + exp\,(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})} \quad (4)$$

logit dari $\pi(\boldsymbol{x_i})$ yaitu

$$g(\boldsymbol{x_i}) = ln\left[\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}\right] \quad (5)$$

sehingga

$$\pi(\boldsymbol{x_i}) = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}} \quad (6)$$

Metode Maksimum Likelihood dapat digunakan untuk mengestimasi parameter pada Regresi Logistik Biner (Agresti, 2007). Setiap pengamatan sebanyak n diasumsikan independen, sehingga fungsi likelihoodnya dapat dihitung sebagai hasil perkalian dari distribusi peluang pada masing-masing pengamatan (Hosmer dan Lemeshow, 2000).

1. Uji Rasio Likelihood (Uji Serentak)

Signifikansi koefisien β terhadap variabel dependen secara bersama-sama dapat diketahui dengan melakukan Uji Rasio Likelihood (Hosmer dan Lemeshow, 2000).
Hipotesis yang diajukan adalah:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ (Seluruh variabel independen tidak ada yang berpengaruh terhadap variabel dependen)

H_1: Paling sedikit ada satu $\beta_j \neq 0$ dengan $j = 1,2, \ldots, p$ (Paling sedikit ada satu variabel independen yang berpengaruh terhadap variabel dependen)

Statistik Uji:

$$G = -2\ln\left[\frac{likelihood\ tanpa\ variabel\ independen}{likelihood\ dengan\ variabel\ independen}\right] \quad (7)$$

H_0 ditolak pada taraf signifikansi α, jika $G > \chi^2_{(\alpha,p)}$ atau $p - value < \alpha$.

2. Uji Wald (Uji Parsial)

Tujuan dari Uji Wald adalah untuk mengetahui pengaruh dari setiap koefisien β_j secara individual dengan membandingkannya terhadap standar error, sehingga dapat diketahui variabel independenddalam model berpengaruh secara signifikan atau tidak terhadap variabel dependennya (Hosmer dan Lemeshow, 2000).
Hipotesis yang digunakan adalah:

H_0: $\beta_j = 0$ dengan $j = 1,2, \ldots, p$ (Tidak ada pengaruh antara variabel independen ke-j dengan variabel dependen)

H_1: $\beta_j \neq 0$ dengan $j = 1,2, \ldots, p$ (Ada pengaruh antara variabel independen ke-j dengan variabel dependen)

Statistik Uji:

$$W_j = \left\{\frac{\widehat{\beta_j}}{SE(\widehat{\beta_j})}\right\} \quad (8)$$

H_0 ditolak pada taraf signifikansi α, jika $W_j > Z_{\alpha/2}$ atau $p - value < \alpha$.

3. Uji Hosmer dan Lemeshow (Uji Kesesuaian Model)

Uji kesesuaian model adalah prosedur yang digunakan untuk mengetahui ada atau tidaknya perbedaan antara prediksi yang dihasilkan oleh model dan hasil observasi sebenarnya dengan kata lain apakah model sesuai atau tidak (Hosmer dan Lemeshow, 2000).

Hipotesis yang digunakan adalah:

H_0: Model sesuai (Tidak ada perbedaan antara prediksi dengan hasil observasi)

H_1: Model tidak sesuai (Ada perbedaan antara prediksi dengan hasil observasi)

Statistik Uji:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{(n'_k \bar{\pi}_k)(1 - \bar{\pi}_k)} \quad (9)$$

H_0 ditolak pada taraf signifikansi α, jika $\hat{C} > \chi^2_{(\alpha,g-2)}$ atau $p - value < \alpha$

CART atau singkatan dari *Classification and Regression Tree* adalah salah satu teknik yang termasuk dalam pohon keputusan (*decision tree*) yang awalnya dikembangkan oleh Breiman, Freidman, Olshen, dan Stone pada tahun 1984. Metode CART melibatkan pembentukan pohon keputusan menggunakan algoritma penyekatan rekursif secara biner (*binery recursive partitioning*). Pohon klasifikasi akan dihasilkan dalam metode CART apabila variabel dependen mempunyai skala kategorik, sedangkan jika variabel dependen berskala kontinu, metode CART akan membentuk pohon regresi (Lewis, 2000). Dasar untuk melakukan proses pemilahan simpul induk menggunakan kriteria pemilahan terbaik (*Goodness of Split Criterion*). Pemilihan pemilah optimal bertujuan untuk mengetahui ukuran tingkat heterogenitas suatu kelas pada simpul tertentu yang dapat dilakukan menggunakan fungsi *impurity measure i(t)* yaitu fungsi Indeks Gini pada persamaan (10).

$$i(t) = \sum_{\substack{j=1 \\ j \neq k}}^{J} P(j|t)P(k|t) \quad (10)$$

Evaluasi pemilihan pemilah *s* pada simpul *t* dapat menggunakan kriteria *goodness of split* sebagai penurunan keheterogenan yang dapat ditentukan menggunakan persamaan (11).

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad (11)$$

Pemilah terbaik didasarkan pada pemilah yang memiliki nilai $\Delta i(s,t)$ tertinggi. Pembentukan pohon berhenti apabila simpul *t* tidak dipilah lagi dan dijadikan simpul terminal. Simpul terminal adalah keadaan Ketika suatu simpul *t* mencapai batas akhir yang ditentukan sehingga tidak terdapat penurunan impuritas secara signifikan sehingga proses pemisahan berhenti (Breiman *et al*., 1993).

AdaBoost dikenalkan oleh Freund dan Schapire pada tahun 1999 yang akan memberikan bobot yang lebih tinggi untuk data yang salah diklasifikasikan. Klasifikasi dasar yang digunakan dalam penelitian ini yaitu CART. Langkah-langkah pada algoritma *Adaptive Boosting* menurut Zhu et al. (2009) adalah sebagai berikut:

1. Melakukan inisialisasi bobot awal pada setiap pengamatan yaitu $w_n^0 = \frac{1}{N}$, dengan N adalah banyaknya amatan yang terdapat pada data latih dengan $n = 1,2,\dots,N$.
2. Untuk iterasi (m) dengan $m = 1,2,\dots,M$
   a. Menetapkan fungsi *classifier* $y^{(m+1)}(x)$ pada data latih dengan bobot w_n.
   b. Mencari nilai $err^{(m+1)}$ menggunakan persamaan (12).

$$err^{(m+1)} = \frac{\sum_{n=1}^{N} w_n{}^{(m)} l\left(y^{(m)}(x_n) \neq t_n\right)}{\sum_{n=1}^{N} w_n{}^{(m)}} \quad (12)$$

   dengan $err^{(m+1)}$ merupakan nilai error pada iterasi ke-*m+1*, $l\left(y^{(m)}(x_n) \neq t_n\right)$ yaitu fungsi indikator yang akan bernilai 1 ketika kelas prediksi $y^{(m)}(x_n)$ tidak sama dengan kelas asli t_n dan bernilai 0 untuk lainnya.

c. Jika nilai $err^{(m+1)} > 1 - \frac{1}{k}$ dengan k adalah banyaknya kelas pada data tersebut, maka iterasi dihentikan. Apabila nilai $err^{(m+1)} \leq 1 - \frac{1}{k}$, selanjutnya dilakukan perhitungan pembobot klasifikasi dengan rumus pada persamaan (13).

$$\alpha^{(m+1)} = \ln\left[\frac{\left(1 - err^{(m+1)}\right)}{err^{(m+1)}}\right] \quad (13)$$

d. Memperbarui nilai bobot amatan menggunakan persamaan (14).

$${w_n}^{(m+1)} = \frac{{w_n}^{(m)}}{\sum_{n=1}^{N} {w_n}^{(m)}} \exp\left(\alpha^{(m+1)} l\left(y^{(m)}(x_n) \neq t_n\right)\right) \quad (14)$$

3. Output

$$T(x) = \arg\max_{k} \sum_{m=0}^{M} \alpha^{(m+1)} l\left(y^{(m+1)}(x_n) = k\right) \quad (15)$$

dengan k merupakan kelas prediksi data yang diuji.

Matriks konfusi diperlukan untuk mengukur kinerja klasifikasi sehingga dapat mengetahui kesalahan klasifikasi yang dihasilkan. Matriks konfusi seperti pada Tabel 1.

Tabel 1. Rumus *Confusion Matrix*

Nilai Sebenarnya	**Nilai Prediksi**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Terdapat beberapa ukuran ketepatan klasifikasi yang dapat digunakan untuk mengukur kebaikan suatu model, antara lain:

$$Akurasi = \frac{TP + TN}{TP + TN + FN + FP} \quad (16)$$

$$Sensitivitas = \frac{TP}{TP + FN} \quad (17)$$

$$Spesifitas = \frac{TN}{TN + FP} \quad (18)$$

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder yang berasal dari UCI (*Univercity of California Irvine*) *Machine Learning Repository*. Data tersebut adalah data Pasien Penyakit Hepatitis C tahun 2020 dan diunduh pada tanggal 25 Oktober 2022. Variabel pada penelitian ini terdiri dari dengan 10 variabel bebas yaitu *Age* (X_1), ALB (X_2), ALP (X_3), ALT (X_4), AST (X_5), BIL (X_6), CHOL (X_7), CREA (X_8), GGT (X_9), PROT (X_{10}) serta 1 variabel respon yaitu *Category* (Y).

Pemrosesan data dilakukan dengan teknik klasifikasi Regresi Logistik Biner dan *Adaptive Boosting* menggunakan software R Studio versi 4.2.2. Langkah-langkah analisis dalam penelitian ini adalah:

1. Melakukan analisis deskriptif terhadap data.

2. Membagi data menjadi data latih dan data uji dengan pendekatan *trial and error* pada beberapa perbandingan, yaitu 50% : 50%, 60% : 40%, 70% : 30%, 80% : 20%, dan 90% : 10%, kemudian diperoleh perbandingan 80% : 20% dengan hasil terbaik.
3. Melakukan pengecekan data *imbalance* pada data latih.
4. Melakukan penanganan data *imbalance* pada data latih menggunakan teknik *Synthetic Minority Over-Sampling Technique* (SMOTE).
5. Melakukan klasifikasi metode Regresi Logistik Biner melalui langkah-langkah sebagai berikut:
   a. Membentuk model Regresi Logistik Biner.
   b. Menentukan estimasi parameter.
   c. Melakukan uji serentak dengan uji Rasio Likelihood.
   d. Melakukan uji parsial dengan uji Wald.
   e. Melakukan uji Hosmer dan Lemeshow untuk mengetahui kesesuaian model pada model Regresi Logistik Biner.
   f. Menentukan model akhir Regresi Logistik Biner.
   g. Menghitung ketepatan klasifikasi model menggunakan data uji.
6. Melakukan klasifikasi menggunakan metode *Adaptive Boosting* dengan tahapan sebagai berikut:
   a. Melakukan transformasi menggunakan normalisasi data.
   b. Melakukan inisialisasi bobot pada data latih $w_n^0 = \frac{1}{N}$
   c. Mencari nilai prediksi data latih menggunakan CART.
   d. Menghitung $err^{(m+1)} = \frac{\sum_{n=1}^{N} w_n{}^{(m)} l\left(y^{(m)}(x_n) \neq t_n\right)}{\sum_{n=1}^{N} w_n{}^{(m)}}$ dengan $l\left(y^{(m)}(x_n) \neq t_n\right)$ yaitu fungsi indikator yang memiliki nilai 1 ketika data amatan *misclassified* dan bernilai 0 untuk lainnya.
   e. Apabila $err^{(m+1)} > 0{,}5$ maka iterasi berhenti. Jika nilai $err^{(m+1)} \leq 0{,}5$ maka dilanjutkan perhitungaan pembobot klasifikasi dengan rumus:
   $\alpha^{(m+1)} = \ln\left[\frac{\left(1-err^{(m+1)}\right)}{err^{(m+1)}}\right]$
   f. Memperbarui nilai bobot baru $w_n{}^{(m+1)} = \frac{w_n{}^{(m)}}{\sum_{n=1}^{N} w_n{}^{(m)}} \exp\left(\alpha^{(m+1)} l\left(y^{(m)}(x_n) \neq t_n\right)\right)$
   g. Menentukan prediksi kelas menggunakan fungsi klasifikasi akhir
   $T(x) = \arg\max_k \sum_{m=0}^{M} \alpha^{(m+1)} l\left(y^{(m+1)}(x_n) = k\right)$
   h. Menghitung ketepatan hasil klasifikasi menggunakan data uji.
7. Membandingkan ketepatan hasil klasifikasi Regresi Logistik Biner dan *Adaptive Boosting* untuk memilih model terbaik.

4. HASIL DAN PEMBAHASAN

Data yang digunakan sebanyak 589 amatan yang terbagi menjadi 533 pasien negatif hepatitis C atau sebesar 90% dan 56 pasien positif hepatitis C atau sebesar 10% dari jumlah keseluruhan amatan pasien. Proporsi terbaik untuk pembagian data dalam penelitian ini adalah 80:20, yaitu 80% merupakan data latih sedangkan sisanya 20% merupakan data uji yang didasarkan pada percobaan beberapa proporsi data. Banyaknya kelas kategori status hepatitis C pasien pada data latih setelah dilakukan SMOTE menjadi seimbang dengan 427 amatan untuk negatif hepatitis dan 427 untuk positif hepatitis.

Model awal Regresi Logistik Biner dibentuk menggunakan estimasi parameter yaitu $\pi(\boldsymbol{x_i}) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$ sebagai peluang positif hepatitis dengan

$$g(\boldsymbol{x_i}) = -48{,}29636 + 0{,}01826X_1 - 0{,}08030X_2 - 0{,}32287X_3 - 0{,}42076\,X_4 + 0{,}47175X_5 + 0{,}15004X_6 - 2{,}85123X_7 + 0{,}09438X_8 + 0{,}13333X_9 + 0{,}81308X_{10}$$

Hipotesis uji rasio Likelihood:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_{10} = 0$ (Seluruh variabel independen tidak ada yang berpengaruh terhadap variabel *category* hepatitis C)

H_1: Paling sedikit ada satu $\beta_j \neq 0$ dengan $j = 1{,}2, \ldots, 10$ (Paling sedikit ada satu variabel independen yang berpengaruh terhadap variabel *category* hepatitis C)

Statistik Uji:

$$G = -2\ln\left[\frac{\text{likelihood tanpa variabel independen}}{\text{likelihood dengan variabel independen}}\right] = 1139{,}2233540$$

H_0 ditolak pada taraf signifikansi $\alpha = 5\%$, karena $G = 1139{,}2233540 > \chi^2_{(0.05,10)} = 18{,}30704$ sehingga dapat disimpulkan bahwa paling sedikit ada satu variabel independen yang berpengaruh terhadap variabel *category* hepatitis C.

Hipotesis:

H_0: $\beta_j = 0$ (Tidak ada pengaruh antara variabel independen ke-j dengan variabel *category* hepatitis C)

H_1: $\beta_j \neq 0$ dengan $j = 1{,}2, \ldots, 10$ (Ada pengaruh antara variabel independen ke-j dengan variabel *category* hepatitis C)

Statistik Uji:

$$W_j = \left\{\frac{\widehat{\beta_J}}{SE(\widehat{\beta_J})}\right\}$$

Hasil uji Wald disajikan seperti pada Tabel 2.

Tabel 2. Hasil Uji Wald Model Awal

Variabel	Estimate	Std.Error	Nilai z	p-value	Keterangan
X_1	0,01826	0,05531	0,330	0,741329	Tidak Signifikan
X_2	-0,08030	0,10734	-0,748	0,454395	Tidak Signifikan
X_3	-0,32287	0,07598	-4,249	2,14e-05	Signifikan
X_4	-0,42076	0,09028	-4,660	3,15e-06	Signifikan
X_5	0,47175	0,11580	4,074	4,63e-05	Signifikan
X_6	0,15004	0,06460	2,323	0,020200	Signifikan
X_7	-2,85123	1,07604	-2,650	0,008055	Signifikan
X_8	0,09438	0,02300	4,104	4,07e-05	Signifikan
X_9	0,13333	0,03197	4,170	3,04e-05	Signifikan
X_{10}	0,81308	0,23977	3,391	0,000696	Signifikan
Constant	-48,29636	14,54767	-3,320	0,000901	-

H_0 ditolak pada taraf signifikansi $\alpha = 5\%$, jika $W_j > Z_{\alpha/2} = 1{,}96$ atau $p - value < \alpha$. Variabel independen yang memiliki pengaruh signifikan terhadap variabel *category* hepatitis C adalah variabel X_3 (ALP), X_4(ALT), X_5 (AST), X_6(BIL), X_7 (CHOL), X_8 (CREA), X_9 (GGT), dan X_{10} (PROT).

Model kedua Regresi Logistik Biner yang terbentuk yaitu $\pi(\boldsymbol{x_i}) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$ sebagai peluang positif hepatitis dengan $g(\boldsymbol{x_i}) = -45{,}23241 - 0{,}30748X_3 - 0{,}43379X_4 + 0{,}47210X_5 + 0{,}12734X_6 - 3{,}09950X_7 + 0{,}08898X_8 + 0{,}12904X_9 + 0{,}75568X_{10}$.

Regresi Logistik Biner menghasilkan matriks konfusi seperti pada Tabel 3 yang mencerminkan tingkat ketepatan klasifikasi dari model tersebut.

Tabel 3. Matriks Konfusi Regresi Logistik Biner

Observasi	Prediksi	
	Negatif Hepatitis (0)	Positif Hepatitis (1)
Negatif Hepatitis (0)	103	3
Positif Hepatitis (1)	0	11

$$Akurasi = \frac{103 + 11}{103 + 11 + 3 + 0} = 0{,}9744$$

$$Sensitivitas = \frac{11}{11 + 0} = 1$$

$$Spesifitas = \frac{103}{103 + 3} = 0{,}97170$$

Ketepatan klasifikasi yang diperoleh menunjukkan bahwa data uji berjumlah 117 data terklasifikasi secara tepat sebanyak 97,44% dan terklasifikasi tidak tepat sebanyak 2,56%. Nilai *Sensitivitas* sebesar 1 sehingga dapat disimpulkan bahwa kelas positif hepatitis yang berlabel 1 diprediksi secara benar sebesar 100%. Nilai *Spesifisitas* sebesar 0,97170 sehingga dapat diketahui bahwa kelas negatif hepatitis yang berlabel 0 diprediksi secara benar sebesar 97,17%. Proses normalisasi telah dilakukan sebelum proses pengolahan data pada klasifikasi *Adaptive Boosting*. Perhitungan bobot awal pada masing-masing data latih adalah $w_n^0 = \frac{1}{N} = \frac{1}{854} = 0{,}00117096$.

Langkah selanjutnya yaitu mencari nilai prediksi data latih menggunakan CART dan batas maksimum iterasi. Perhitungan CART dengan mencari indeks gini pada variabel dependen dan mencari nilai *goodness of split* pada setiap variabel independen. Nilai *goodness of split* tertinggi dijadikan sebagai pemilah. Prediksi data latih yang telah diperoleh menggunakan CART, dilanjutkan dengan mencari nilai $err^{(1)}$ menggunakan Persamaan 15. Jika nilai $err^{(1)} < 0{,}5$ perhitungan dilanjutkan dengan mencari $\alpha^{(1)}$ dengan rumus pada Persamaan 18. Nilai $\alpha^{(1)}$ telah diperoleh selanjutnya memperbarui bobot dengan menggunakan Persamaan 14. Perhitungan berulang dan berhenti jika nilai $err^{(m+1)} > 0{,}5$ atau perhitungan sampai iterasi maksimum. Prediksi kelas ditentukan dengan melihat jumlah α tertinggi pada masing-masing kelas.

Ketepatan klasifikasi dengan *Adaptive Boosting* menghasilkan matriks konfusi seperti pada Tabel 4.

Tabel 4. Matriks Konfusi Adaptive *Boosting*

Observasi	Prediksi	
	Negatif Hepatitis (0)	Positif Hepatitis (1)
Negatif Hepatitis (0)	104	1
Positif Hepatitis (1)	2	10

$$Akurasi = \frac{10 + 104}{10 + 104 + 1 + 2} = 0{,}9744$$

$$Sensitivitas = \frac{10}{10 + 1} = 0{,}9091$$

$$Spesifitas = \frac{104}{104 + 2} = 0{,}9811$$

Ketepatan klasifikasi yang diperoleh menunjukkan bahwa data uji berjumlah 117 data terklasifikasi secara tepat sebanyak 97,44% dan terklasifikasi tidak tepat sebanyak 2,56%. Nilai *Sensitivitas* sebesar 0,9091 sehingga dapat disimpulkan bahwa kelas positif hepatitis yang berlabel 1 diprediksi secara benar sebesar 90,91%. Nilai *Spesifisitas* sebesar 0,9811 sehingga dapat diketahui bahwa kelas negatif hepatitis yang berlabel 0 diprediksi secara benar sebesar 98,11%.

5. KESIMPULAN

Hasil ketepatan klasifikasi menunjukkan bahwa metode Regresi Logistik Biner mencapai tingkat akurasi sebesar 97,44% dengan nilai *sensitivitas* 100% dan *spesifisitas* 97,17%. Klasifikasi menggunakan metode CART dengan *Adaptive Boosting* dapat diterapkan dan menghasilkan nilai akurasi sebesar 97,44% dengan nilai *sensitivitas* 90,91% dan *spesifisitas* 98,11%. Metode terbaik yang dapat mengklasifikasi dan memprediksi status hepatitis C pasien dengan baik pada penelitian ini adalah Regresi Logistik Biner karena ketepatan hasil klasifikasi Regresi Logistik Biner lebih tinggi dari pada *Adaptive Boosting* yaitu dengan nilai sensitivitas sebesar 100%

DAFTAR PUSTAKA

Agresti, A. 2007. *An Introduction to Categorical Data Analysis Second Edition*. Florida: Wiley.

Ali, A., Shamsuddin, S. M., dan Ralescu, A. L. 2015. Classification with Class Imbalance Problem. *Int. J. Advance Soft Compu. Appl* Vol. 5, No. 3.

Breiman L., Friedman J.H., Olshen R.A., dan Stone C.J. 1993. *Classification And Regression Trees*. New York: Chapman And Hall

Chawla, N. V., Bowyer, K. W., Hall, L. O., dan Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, Vol 16, Hal: 321-357.

Hosmer, D. W., dan Lemeshow, S. 2000. *Apllied Logistic Regression*. New York: John Wiley & Sons.

Ikatan Dokter Anak Indonesia. 2009. *Buku Ajar Gastroenterologi-Hepatologi Jilid 1*. UKK – Gastroenterologi-Hepatologi IDAI.

Lewis, R. J. 2000. An Introduction to Classification and Regression Tree (CART) Analysis. *Proceedings of Annual Meeting of the Society for Academic Emergency Medicine*, San Francisco, CA, USA.

Perhimpunan Peneliti Hati Indonesia, 2017. *Konsesnsus Nasional Penatalaksanaan Hepatitis C di Indonesia*. Jakarta: PPHI.

Prasetyo, E. 2012. *DATA MINING : Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta : ANDI

Qiong, GU, et al. 2016. An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification. *J Dig Inf Manag* Vol. 14, No. 2, Hal: 92–103.

Rabbani, R., Wahidah, I., dan Santoso, I. H. 2021. Klasifikasi Data Deteksi Jatuh Menggunakan Machine Learning dengan Algoritma *Adaptive Boosting* (*Adaboost*). *e-Proceeding of Engineering* Vol.8, No.5.

Suwardika, G. 2017. Pengelompokan dan Klasifikasi pada Data Hepatitis dengan Menggunakan Support Vector Machine (SVM), Classification and Regression Tree

(Cart) dan Regresi Logistik Biner. *Journal of Education Research and Evaluation* Vol. 1, No. 3, Hal: 183-191.

World Health Organization. 2022. Hari Hepatitis Dunia. Tersedia: https://www.who.int/indonesia/news/campaign/world-hepatitis-day (diakses pada tanggal 10 November 2022).

Zhu, J., Zou, H., Rosset, S., dan Hastie, T. 2009. Multi-class AdaBoost. *Statistics and Its Interface* Vol.2, Hal: 340-360.