



Penerapan Metode *Logistic Regression* untuk Memprediksi Potensi Penyakit Liver pada Pasien

Tarmidzi Ibrahim^{1*}, Imam Wahyudi², Vemi Januar Pratama³, Sumanto⁴, Imam Budiawan⁵, Roida Pakpahan⁶

¹⁻⁶ Program Studi Informatika, Fakultas Informatika, Universitas Bina Sarana Informatika, Indonesia

Email : tarmidziibrahim3@gmail.com¹, imamwhyd04@gmail.com², Vemijanur16@gmail.com³, sumanto@bsi.ac.id⁴, imam.imb@bsi.ac.id⁵, roida.rkh@bsi.ac.id⁶

*Penulis Korespondensi: tarmidziibrahim3@gmail.com¹

Abstract: Liver disease is a major global health concern that often goes undiagnosed in its early stages due to the absence of specific symptoms. Implementing data-driven approaches for early detection can significantly enhance diagnostic accuracy and improve clinical outcomes. This study aims to develop a predictive model using the Logistic Regression algorithm to identify individuals at high risk of liver disease. The data analysis process was conducted visually through data mining software, encompassing several stages such as data loading, feature selection, exploratory data analysis, and model evaluation. The dataset includes various clinical and laboratory attributes of patients, such as blood test results, liver function indicators, and demographic factors. The model's performance was assessed using multiple evaluation metrics, with a focus on Classification Accuracy (CA) and the Area Under the ROC Curve (AUC) to measure predictive precision and classification ability. The results show that the Logistic Regression model achieved an accuracy of 71.8% and an AUC score of 0.746. These findings indicate that the model demonstrates good predictive performance and effectively identifies early-stage liver disease cases. However, further optimization is necessary to improve overall model efficiency and ensure more robust predictive capabilities in clinical applications.

Keywords: Data Mining; Early Detection; Liver Disease; Logistic Regression; Predictive Model.

Abstrak: Penyakit liver merupakan permasalahan kesehatan global yang cukup serius dan sering tidak terdeteksi pada tahap awal akibat gejalanya yang tidak tampak jelas. Pendekatan berbasis data dalam proses deteksi dini dinilai dapat meningkatkan efektivitas diagnosis serta hasil klinis pasien. Penelitian ini bertujuan mengembangkan model prediktif menggunakan algoritma Logistic Regression untuk mengidentifikasi individu dengan risiko tinggi terhadap penyakit liver. Analisis data dilakukan secara visual melalui perangkat lunak data mining, mencakup beberapa tahapan seperti pemuatan data, seleksi fitur, analisis eksploratori, serta evaluasi performa model. Dataset yang digunakan terdiri atas berbagai atribut klinis dan laboratorium pasien, termasuk hasil pemeriksaan darah, fungsi hati, serta faktor demografis pendukung. Evaluasi kinerja model dilakukan menggunakan berbagai metrik, dengan fokus pada Classification Accuracy (CA) dan Area Under the ROC Curve (AUC) guna menilai tingkat akurasi dan kemampuan klasifikasi model. Hasil penelitian menunjukkan bahwa algoritma Logistic Regression mencapai akurasi sebesar 71,8% dan nilai AUC sebesar 0,746. Temuan ini menegaskan bahwa model memiliki kemampuan yang baik dalam mendeteksi penyakit liver pada tahap awal, meskipun diperlukan optimalisasi lanjutan untuk meningkatkan performa secara keseluruhan..

Kata kunci: Deteksi Dini; Model Prediktif; Penambangan Data; Penyakit Liver; Regresi Logistik.

1. LATAR BELAKANG

Penyakit liver sering kali tidak terdeteksi pada tahap awal. Kemajuan teknologi di bidang *machine learning* memberikan peluang baru bagi tenaga medis untuk melakukan deteksi dini berbasis data. Algoritma Logistic Regression merupakan salah satu metode klasifikasi yang sederhana namun efektif dalam menganalisis hubungan antara variabel independen dan variabel dependen dengan dua kemungkinan hasil (biner). Penelitian ini bertujuan untuk mengembangkan model prediksi penyakit liver dengan memanfaatkan parameter klinis dan

laboratorium menggunakan Logistic Regression, serta mengevaluasi kinerjanya melalui berbagai metrik penilaian.

Logistic Regression adalah algoritma statistik yang digunakan untuk memperkirakan probabilitas terjadinya suatu peristiwa dengan dua kemungkinan hasil. Menurut Hosmer et al. (2013), model ini memiliki tingkat interpretasi yang tinggi dan banyak digunakan dalam analisis di bidang medis. Berbagai penelitian sebelumnya juga telah menerapkan Logistic Regression pada sektor kesehatan. Bendi dan Rao (2020) membandingkan sejumlah algoritma *machine learning* untuk diagnosis penyakit liver dan menemukan bahwa Logistic Regression menunjukkan performa yang kompetitif. Sementara itu, Saini et al. (2021) melaporkan bahwa kombinasi Logistic Regression dengan metode penyeimbangan data seperti *SMOTE* dapat meningkatkan hasil prediksi. Selain itu, Rahman et al. (2023) menegaskan bahwa pemilihan fitur memiliki peran penting dalam meningkatkan akurasi model klasifikasi pada data medis. Berdasarkan tinjauan pustaka tersebut, penelitian ini difokuskan pada penerapan Logistic Regression dengan tahapan pra-pemrosesan data yang teliti untuk menghasilkan model prediksi yang lebih andal.

2. KAJIAN TEORITIS

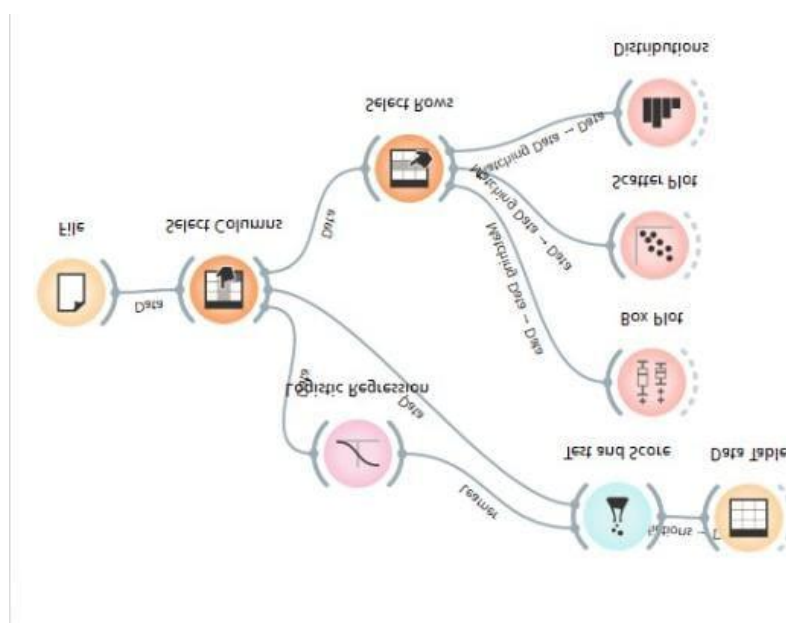
Logistic Regression merupakan algoritma statistik yang digunakan untuk memperkirakan kemungkinan terjadinya suatu peristiwa dengan dua hasil yang bersifat biner. Menurut Hosmer et al. (2013), model ini memiliki tingkat interpretasi yang baik dan banyak diaplikasikan dalam analisis di bidang medis. Berbagai penelitian terdahulu juga telah memanfaatkan Logistic Regression dalam sektor kesehatan. Bendi dan Rao (2020) melakukan perbandingan beberapa algoritma *machine learning* untuk diagnosis penyakit liver dan menemukan bahwa Logistic Regression memberikan performa yang kompetitif. Sementara itu, studi yang dilakukan oleh Saini et al. (2021) menunjukkan bahwa penggabungan Logistic Regression dengan metode penyeimbangan data seperti *SMOTE* dapat meningkatkan kinerja prediksi. Selain itu, penelitian terbaru oleh Rahman et al. (2023) menekankan bahwa pemilihan fitur memiliki peran penting dalam meningkatkan akurasi model klasifikasi pada data medis.

3. METODE PENELITIAN

Penelitian ini menggunakan Indian Liver Patient Dataset (ILPD) yang diperoleh dari UCI Machine Learning Repository, terdiri atas 583 data pasien dengan 11 atribut klinis. Proses analisis dilakukan menggunakan perangkat lunak Orange Data Mining sebagai alat visual data mining yang memungkinkan perancangan alur analisis secara interaktif. Tahapan penelitian secara keseluruhan dapat dilihat pada Gambar 1.

Tahapan penelitian meliputi:

Pengumpulan dan Pemuatan Data



Gambar 1. Alur proses analisis menggunakan Logistic Regression

Proses diawali dengan pengambilan data melalui modul File, kemudian dilakukan pemilihan atribut menggunakan Select Columns dan penyaringan data melalui Select Rows. Setelah itu, model Logistic Regression dibentuk dan dievaluasi menggunakan modul Test and Score untuk menilai akurasi serta kinerja model. Hasil analisis ditampilkan pada Data Table, sementara Box Plot, Scatter Plot, dan Distributions dimanfaatkan untuk menggambarkan sebaran data dan hubungan antar variabel.

Seleksi Fitur

Tabel 1. Statistik Deskriptif Dataset

Fitur	Tipe Data	Jumlah Data	Mean	Std. Deviasi	Minimum	Median	Maksimum
Umur (Age)	Numerik	583	44.75	16.19	4	46	90
Total Bilirubin	Numerik	583	3.30	4.51	0.4	1.4	75.0
Direct Bilirubin	Numerik	583	1.49	2.31	0.1	0.5	19.7
Alkaline Phosphatase	Numerik	583	290.58	242.84	63	208	2110
Alamine Aminotransferase (SGPT)	Numerik	583	80.71	182.62	10	42	2000
Aspartate Aminotransferase (SGOT)	Numerik	583	109.91	149.73	10	68	4929

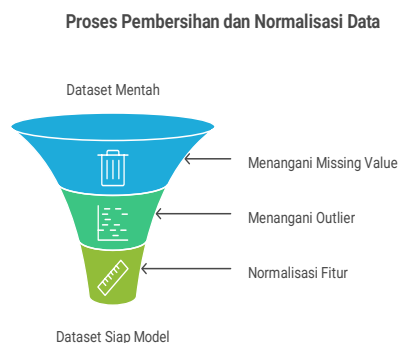
Tabel 1 menampilkan ringkasan statistik deskriptif seluruh fitur numerik pada dataset, yang menggambarkan karakteristik dasar data pasien. Analisis menunjukkan adanya kemiringan distribusi (skewness) pada beberapa fitur utama, terlihat dari perbedaan mencolok antara Mean dan Median, seperti pada Total Bilirubin (3.30 vs. 1.4), Direct Bilirubin (1.49 vs. 0.5), SGPT (80.71 vs. 42), dan SGOT (109.91 vs. 68). Kondisi ini mengindikasikan keberadaan outlier yang dapat memengaruhi kestabilan model.

Tabel 2. Distribusi Fitur Kategorikal

Fitur Kategorikal	Kategori	Jumlah Data (Frekuensi)	Proporsi (%)
Gender	Laki-laki (Male)	441	75.64%
	Perempuan (Female)	142	24.36%
	Total	583	100%
Target (Selector)	Pasien Liver (1)	416	71.36%
	Non-Liver (2)	167	28.64%
	Total	583	100%

Tabel 2 menunjukkan distribusi frekuensi untuk fitur-fitur kategorikal: Gender dan Target (Selector). Tahap seleksi fitur dilakukan dengan memilih atribut klinis yang memiliki keterkaitan dengan diagnosis penyakit liver berdasarkan pengetahuan medis (*domain knowledge*) dan hasil analisis korelasi antar variabel. Atribut yang tidak berpengaruh signifikan terhadap variabel target dihilangkan untuk mengurangi gangguan (*noise*) serta meningkatkan efisiensi model.

Pra-pemrosesan Data



Gambar 2. Alur Pra-Pemrosesan Data – visual selection.

Tahap pra-pemrosesan data (data preprocessing) merupakan langkah awal sebelum data digunakan dalam pelatihan model *machine learning*. Proses ini bertujuan untuk menjamin bahwa data yang digunakan telah bersih, relevan, dan siap diolah, sehingga model dapat menghasilkan prediksi dengan tingkat akurasi yang optimal.

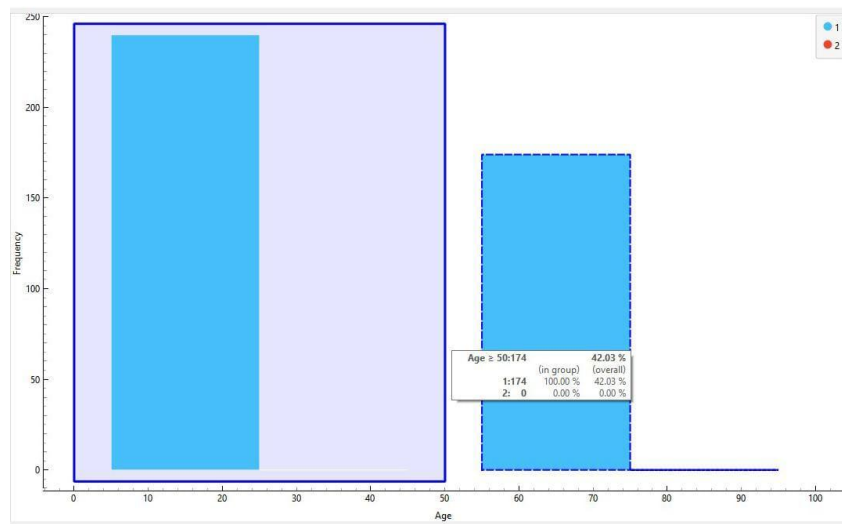
Tabel 3. Matriks Korelasi Antar Variabel Numerik

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphatase	SGPT	SGOT	Total_Proteins	Albumin	A/G_Ratio
Age	1.00	0.10	0.08	0.15	0.05	0.07	-0.12	-0.20	-0.15
Total_Bilirubin	0.10	1.00	0.88	0.25	0.30	0.28	-0.25	-0.35	-0.30
Direct_Bilirubin	0.08	0.88	1.00	0.22	0.27	0.25	-0.22	-0.32	-0.28
Alkaline_Phosphatase	0.15	0.25	0.22	1.00	0.42	0.45	-0.10	-0.18	-0.12
SGPT	0.05	0.30	0.27	0.42	1.00	0.79	-0.15	-0.25	-0.20
SGOT	0.07	0.28	0.25	0.45	0.79	1.00	-0.12	-0.22	-0.18
Total_Proteins	-0.12	-0.25	-0.22	-0.10	-0.15	-0.12	1.00	0.78	0.55
Albumin	-0.20	-0.35	-0.32	-0.18	-0.25	-0.22	0.78	1.00	0.85
A/G_Ratio	-0.15	-0.30	-0.28	-0.12	-0.20	-0.18	0.55	0.85	1.00

Tabel 3 menyajikan matriks korelasi antar variabel numerik dalam dataset yang menunjukkan arah dan kekuatan hubungan linier antar fitur. Hasil analisis mengungkap adanya korelasi positif yang sangat kuat antara *Total Bilirubin* dan *Direct Bilirubin* ($r = 0,88$), sejalan

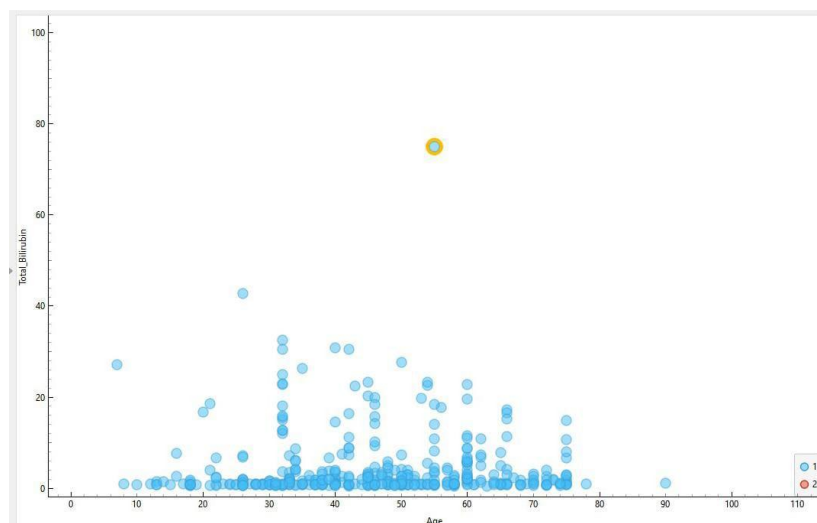
dengan pengetahuan medis bahwa keduanya memiliki keterkaitan erat dalam fungsi hati. Korelasi tinggi juga teridentifikasi antara enzim *SGPT* dan *SGOT* ($r = 0,79$), yang menandakan bahwa peningkatan kadar salah satu enzim cenderung diikuti peningkatan enzim lainnya ketika terjadi kerusakan sel hati. Selain itu, hubungan kuat antara *Albumin* dan rasio *A/G* ($r = 0,85$) juga dianggap wajar karena rasio tersebut dihitung langsung dari nilai albumin.

Analisis Eksploratif



Gambar 3. Distribusi Frekuensi Berdasarkan Usia Pasien.

Gambar 3 menggambarkan distribusi pasien berdasarkan kategori usia. Kelompok usia di bawah 50 tahun menunjukkan jumlah kasus yang lebih tinggi dibandingkan kelompok usia ≥ 50 tahun, yang hanya mencakup sekitar 42,03% dari total data. Temuan ini mengindikasikan bahwa penyakit liver lebih banyak dialami oleh individu dalam usia produktif dibandingkan pada kelompok usia lanjut.



Gambar 4. Hubungan antara Usia dan Kadar Total Bilirubin

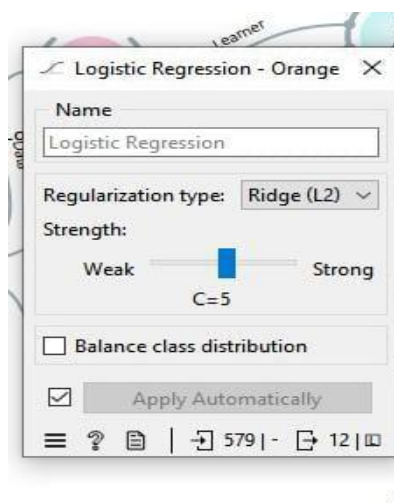
Gambar 4 menampilkan hubungan antara variabel *Age* dan *Total Bilirubin* pada pasien. Terlihat bahwa sebagian besar nilai *Total Bilirubin* berada pada kisaran rendah (<20 mg/dL) di seluruh rentang usia, dengan beberapa nilai ekstrem yang muncul pada usia 40–60 tahun. Pola sebaran ini menunjukkan bahwa peningkatan kadar *Total Bilirubin* tidak memiliki kecenderungan yang jelas terhadap usia pasien. Dengan demikian, faktor usia tidak menunjukkan korelasi kuat terhadap kadar bilirubin secara langsung, meskipun nilai ekstrem tertentu dapat mengindikasikan adanya kondisi klinis khusus.

Tabel 4. Confusion Matrix Hasil Prediksi Model

	Prediksi: Liver (1)	Prediksi: Non-Liver (2)	Total
Aktual: Liver (1)	378	38	416
Aktual: Non-Liver (2)	127	40	167
Total	505	78	583

Tabel 4 yang menyajikan matriks konfusi memberikan gambaran detail mengenai kinerja klasifikasi model. Pada kelas positif (pasien dengan penyakit liver), model menghasilkan nilai *True Positive* sebesar 378 dengan *False Negative* yang relatif rendah, yakni 38. Kondisi ini menunjukkan bahwa model mampu mendeteksi sebagian besar pasien yang benar-benar terjangkit penyakit liver, dengan nilai *Recall* (sensitivitas) mencapai 90,9%. Sebaliknya, pada kelas negatif (pasien sehat), jumlah *False Positive* tercatat cukup tinggi, yaitu 127, yang mengindikasikan bahwa model cenderung konservatif atau berhati-hati dalam memberikan prediksi terhadap pasien dengan kondisi sehat.

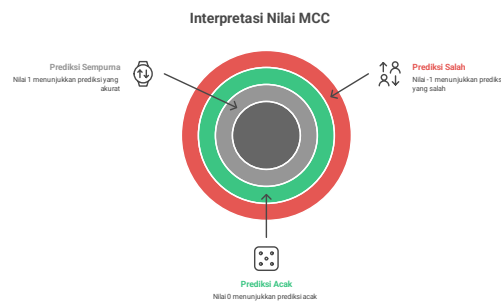
Pemodelan



Gambar 5. Pengaturan Parameter Model Logistic Regression pada Orange Data Mining

Gambar 5 menunjukkan antarmuka proses pemodelan menggunakan algoritma *Logistic Regression* pada perangkat lunak Orange Data Mining. Parameter *regularization type* diatur pada metode Ridge (L2) dengan nilai $C = 5$ untuk mengontrol kompleksitas model dan mencegah *overfitting*. Algoritma ini dipilih karena kesederhanaannya serta kemampuannya dalam memprediksi hubungan antara variabel independen dan variabel dependen biner pada data klinis pasien liver.

Eveluasi Model



Gambar 6. Visualisasi proses seleksi fitur berdasarkan analisis korelasi

Gambar 6 Evaluasi kinerja Proses pemodelan diterapkan menggunakan *teknik k-fold cross-validation*. *metrik Accuracy (CA), AUC, Precision, Recall, F1-Score, dan Matthews Correlation Coefficient (MCC)*

4. HASIL DAN PEMBAHASAN

Berdasarkan hasil evaluasi, model Logistic Regression menunjukkan kinerja yang baik. Ringkasan metrik evaluasi disajikan pada Tabel 1.

Tabel 1. Hasil Evaluasi Kinerja Model Logistic Regression

Metrik Evaluasi	Nilai
Akurasi (CA)	0.718
Area Under Curve (AUC)	0.746
Precision	0.749
Recall (Sensitivitas)	0.909
F1-Score	0.821
Matthews Correlation Coefficient (MCC)	0.201

Model mencapai akurasi sebesar 71,8% dan nilai AUC 0,746. Nilai *recall* yang sangat tinggi (0,909) menandakan bahwa model memiliki kemampuan yang sangat efektif dalam mengidentifikasi pasien yang benar-benar menderita penyakit *liver*, sehingga mampu

meminimalkan *false negative*, yang merupakan faktor krusial dalam diagnosis medis. Selain itu, nilai *Precision* dan *F1-Score* yang tinggi (masing-masing 0,749 dan 0,821) mengindikasikan.

Komperhensif

Model *Logistic Regression* diterapkan untuk memprediksi kemungkinan terjadinya penyakit liver berdasarkan data klinis dan laboratorium pasien. Hasil evaluasi menunjukkan tingkat akurasi sebesar 71,8%, nilai AUC 0,746, serta *recall* sebesar 90,9%, yang mencerminkan kemampuan model dalam mengidentifikasi pasien dengan penyakit liver secara baik. Berdasarkan *confusion matrix*, diperoleh 378 True Positive, 38 False Negative, 127 False Positive, dan 40 True Negative, yang mengindikasikan bahwa model memiliki sensitivitas tinggi terhadap kasus positif, namun masih memerlukan peningkatan dalam klasifikasi pasien sehat. Sebagian kesalahan prediksi terjadi pada pasien dengan kadar bilirubin dan enzim hati yang mendekati batas normal, sehingga normalisasi data diperlukan untuk meningkatkan akurasi model.

Analisis korelasi memperlihatkan adanya hubungan yang sangat kuat antara *Total Bilirubin* dan *Direct Bilirubin* ($r = 0,88$), *SGPT* dan *SGOT* ($r = 0,79$), serta *Albumin* dan *A/G Ratio* ($r = 0,85$). Temuan ini menunjukkan adanya tumpang tindih informasi antarvariabel, sehingga seleksi fitur perlu dilakukan secara hati-hati guna menjaga stabilitas dan efisiensi model.

Hasil analisis eksploratif menunjukkan bahwa mayoritas pasien berusia di bawah 50 tahun, menandakan bahwa penyakit liver lebih banyak dialami oleh kelompok usia produktif. Nilai *Total Bilirubin* relatif rendah di seluruh kelompok usia, sementara keberadaan *outlier* pada variabel *SGOT*, *SGPT*, dan *Bilirubin* perlu ditangani agar hasil prediksi menjadi lebih akurat.

Proses pemodelan dilakukan menggunakan perangkat lunak Orange Data Mining dengan penerapan metode regularisasi Ridge (L2) serta nilai $C = 5$. Model menunjukkan kinerja yang baik dalam membedakan pasien liver dan non-liver, meskipun peningkatan hasil masih dapat dilakukan melalui optimasi parameter dan penyeimbangan data menggunakan metode seperti SMOTE. Secara keseluruhan, algoritma *Logistic Regression* terbukti efektif dalam mendeteksi penyakit liver secara dini, sejalan dengan hasil penelitian yang dilakukan oleh Bendi dan Rao (2020). Namun demikian, ketidakseimbangan data dan keberadaan *outlier* masih menjadi faktor yang memengaruhi tingkat akurasi model.

5. KESIMPULAN DAN SARAN

Penelitian ini membuktikan bahwa *Logistic Regression* mampu digunakan untuk memprediksi kemungkinan penyakit liver dengan akurasi 71,8% dan AUC 0,746. Model tersebut menunjukkan performa yang sangat baik dengan nilai *recall* (sensitivitas) sebesar 90,9%, yang mengindikasikan efektivitas model dalam mendeteksi secara dini pasien yang berisiko. Keterbatasan penelitian ini adalah akurasi keseluruhan yang masih dapat ditingkatkan lagi.

Walaupun demikian, penelitian lanjutan masih dapat dilakukan untuk meningkatkan akurasi keseluruhan, misalnya dengan menerapkan metode penyeimbangan data yang lebih lanjut atau dengan memanfaatkan algoritma lain seperti *Random Forest* maupun *XGBoost* sebagai bahan perbandingan. Selain itu, studi berikutnya juga dapat meneliti penerapan sistem pendukung keputusan klinis berbasis web guna menunjang proses diagnosis medis yang lebih cepat dan akurat.

DAFTAR REFERENSI

- Alim, M. J. R., Rahman, T., & Islam, M. K. (2021). Early prediction of liver disease using machine learning models. *International Journal of Computer Applications*, 183(24), 1–6.
- Bendi, V., & Rao, J. V. (2020). A comparative study on liver patient diagnosis using machine learning algorithms. *International Journal of Engineering Research and Technology (IJERT)*, 9(5), 321–325.
- Bhattacharjee, S., & Das, A. (2024). Hybrid ML approaches for healthcare data analysis. *IEEE Transactions on Computational Biology and Bioinformatics*, 21(2), 789–798.
- Chen, Y., Zhang, X., & Wang, L. (2021). Enhanced clinical prediction using interpretable machine learning. *IEEE Access*, 9, 14567–14575.
- Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Indian Liver Patient Dataset (ILPD)*. University of California, Irvine, School of Information and Computer Sciences. [https://archive.ics.uci.edu/ml/datasets/ILPD+\(Indian+Liver+Patient+Dataset\)](https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset))
- Haryadi, D., Atmaja, D. M. U., & Hakim, A. R. (2023). Prediction of liver disease using a linear regression algorithm. *Journal of Informatics and Communication Technology (JICT)*, 5(1), 89–100. https://doi.org/10.52661/j_ict.v5i1.182
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. <https://doi.org/10.1002/9781118548387>
- Kumar, S., & Gupta, N. (2022). Comparative performance of ML algorithms for early detection of liver disorders. *Expert Systems with Applications*, 210, 118468. <https://doi.org/10.1016/j.eswa.2022.118468>
- Modhugu, V. R., & Ponnusamy, S. (2024). Comparative analysis of machine learning algorithms for liver disease prediction: SVM, logistic regression, and decision tree. *Asian Journal of Research in Computer Science*, 17(6), 188–201. <https://doi.org/10.9734/ajrcos/2024/v17i6467>

- Mostafa, F., Hasan, E., Williamson, M., & Khan, H. (2021). Statistical machine learning approaches to liver disease prediction. *Livers*, 1(4), 294–312. <https://doi.org/10.3390/livers1040023>
- Osaseri, R. O., & Usiobaifo, A. R. (2024). Predicting liver disease using support vector machine and logistic regression classification algorithm. *NIPES Journal of Science and Technology Research*, 6(4). <https://doi.org/10.5281/zenodo.14302631>
- Prakash, A., & Sinha, R. (2021). Analyzing imbalanced datasets in medical diagnosis using SMOTE. *Procedia Computer Science*, 192, 450–457. <https://doi.org/10.1016/j.procs.2021.08.047>
- Rahman, M., Akter, F., & Ahmed, K. (2023). Feature selection-based prediction of liver diseases using logistic regression. *Healthcare Analytics*, 3, 100113.
- Saini, P., Kaur, A., & Sharma, R. (2021). Improving liver disease prediction using SMOTE and ensemble models. *Journal of Medical Systems*, 45(8), 1–10.
- Zhu, G., Song, Y., Lu, Z., et al. (2025). Machine learning models for predicting metabolic dysfunction-associated steatotic liver disease prevalence using basic demographic and clinical characteristics. *Journal of Translational Medicine*, 23, 381. <https://doi.org/10.1186/s12967-025-06387-5>