VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

Analisis Model Klasifikasi untuk Prediksi Penyakit Liver Menggunakan Regresi Logistik dan Random Forest

Puji Triatmo[1], Mohammad Wildan Aghna[2], Pratama Putra Sabila[3]
[123]Program Studi Informatika, Universitas Bina Sarana Informatika, Jakarta
15230187@bsi.ac.id[1], 15230493@bsi.ac.id[2], 15230194@bsi.ac.id[3]

ABSTRACT

Liver disease is a serious healthcondition that requires early detection to prevent further complications. This study aims to evaluate the performance of classification models in predicting liver disease using Logistic Regression and Random Forest algorithms. The dataset was obtained from the Kaggle platform and includes clinical variabels such as age, gender, bilirubin levels, liver enzymes, protein levels, and albumin. The research stages consist of data preprocessing, exploratory data analysis (EDA), data splitting into training and testing sets, and model development using both algorithms. Model performance was evaluated using accuracy, precision, recall, and F1-Score based on the confusion matrix. The results indicate that Logistic Regression achieves a very high recall value of 0.99, making it effective in detecting positive liver disease cases. Meanwhile, Random Forest demonstrates a more stable performance with an accuracy of 0.75 and precision of 0.76. Both models obtain the same F1-Score of 0.84, indicating a balanced performance between precision and recall. Overall, both algorithms provide reliable prediction results, with Logistic Regression being more suitable for early detection and Random Forest offering more stable and balanced predictions.

Keywords: *liver disease, classification, logistic regression, random forest, machine learning*

ABSTRAK

Penyakit liver merupakan salah satu gangguan kesehatan serius yang membutuhkan deteksi dini untuk mencegah komplikasi lebih lanjut. Penelitian ini bertujuan untuk menganalisis performa model klasifikasi dalam memprediksi penyakit liver menggunakan algoritma Regresi Logistik dan Random Forest. Dataset diperoleh dari platform Kaggle dan mencakup variabel klinis seperti usia, jenis kelamin, kadar bilirubin, enzim hati, protein, dan albumin. Tahapan penelitian meliputi pengolahan data, *exploratory data analysis* (EDA), pembagian data menjadi data latih dan data uji, serta pembangunan model menggunakan kedua algoritma. Evaluasi kinerja dilakukan menggunakan metrik accuracy, precision, recall, dan F1-Score berdasarkan confusion matrix. Hasil penelitian menunjukkan bahwa Regresi Logistik memiliki nilai recall yang sangat tinggi sebesar 0,99, sehingga unggul dalam mendeteksi kasus positif penyakit liver. Sementara itu, Random Forest memberikan performa yang lebih stabil dengan akurasi sebesar 0,75 dan precision sebesar 0,76. Kedua model menghasilkan nilai F1-Score yang sama yaitu 0,84, yang menunjukkan keseimbangan performa yang baik. Secara keseluruhan, kedua algoritma mampu memberikan hasil prediksi yang reliabel, dengan Regresi Logistik lebih sesuai untuk deteksi dini dan Random Forest lebih tepat untuk menghasilkan prediksi yang stabil dan seimbang.

Kata kunci: penyakit liver, klasifikasi, regresi logistik, random forest, *machine learning*

PENDAHULUAN

Penyakit liver merupakan satu isu kesehatan yang berdampak signifikan terhadap kualitas hidup dan tingkat mortalitas populasi. Gangguan fungsi hati dapat

Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

memicu komplikasi serius seperti sirosis, hepatitis kronis, serta gangguan metabolik yang memengaruhi kondisi tubuh secara luas. Oleh karena itu, deteksi dini terhadap penyakit hati menjadi langkah krusial dalam upaya pencegahan dan penanganan selanjutnya. Perkembangan teknologi data dan machine learning menawarkan peluang besar dalam pengembangan model prediksi yang dapat mendukung proses diagnosis yang lebih cepat, akurat, dan konsisten. Pendekatan ini memfasilitasi analisis data pasien secara menyeluruh melalui identifikasi pola dan keterkaitan antarvariabel klinis dengan lebih tepat dibandingkan metode konvensional.

Beberapa penelitian sebelumnya telah menerapkan berbagai metode *machine learning* dalam memprediksi penyakit liver. (Hikmah, 2025) mengembangkan model berbasis metode ensemble yang menunjukkan peningkatan performa prediksi melalui optimasi hyperparameter. (Mardewi, 2023) menggunakan beberapa algoritma seperti Logistik Regression, Decision Tree, dan SVM dalam mendeteksi dini penyakit liver dan menunjukkan bahwa model *machine learning* mampu meningkatkan akurasi diagnosis. Selain itu, (Sen et al., 2025) mengkaji performa beberapa algoritma termasuk Logistic Regression dan Random Forest pada dataset ILPD dan menemukan bahwa algoritma ensemble menunjukkan hasil yang lebih stabil dalam klasifikasi.

Penelitian internasional terkini juga menunjukkan bahwa pendekatan *machine learning* memiliki potensi besar dalam diagnosis penyakit liver. (Zhang et al., 2025) mengembangkan model prediksi risiko *non-alcoholic fatty liver disease* (NAFLD) menggunakan berbagai algoritma machine learning dan membuktikan bahwa model prediktif dapat membantu pengambilan keputusan klinis. Penelitian oleh (Herliawan et al., 2020) juga mengaplikasikan Random Forest dan *feature selection* backward elimination untuk meningkatkan akurasi klasifikasi penyakit liver. Sementara itu, (Lu, 2023) menegaskan bahwa kombinasi beberapa algoritma *machine learning* mampu memberikan hasil prediksi yang lebih optimal.

Kajian lainnya juga memperlihatkan tren penggunaan algoritma klasifikasi dalam diagnosis medis. (Nurkholifah et al., 2023) membandingkan performa beberapa algoritma dalam memprediksi penyakit liver dan menemukan bahwa model sangat dipengaruhi oleh kualitas data serta teknik pra-pemrosesan. Penelitian terbaru oleh (An et al., 2025) mengembangkan model prediksi *Metabolic Dysfunction-Associated Steatotic Liver Disease* menggunakan metode seperti Logistic Regression, Random Forest, dan XGBoost. Selain itu, penelitian terkait dimensionality reduction oleh (Karna et al., 2024) menunjukkan bahwa teknik reduksi dan pemilihan fitur mampu meningkatkan performa model klasifikasi. Beberapa penelitian juga menggabungkan pendekatan algoritma Logistic Regression dan Random Forest dalam domain medis lainnya, seperti prediksi penyakit jantung (Silmi Ath Thahirah Al Azhima1, 2022), yang mendukung kesesuaian kedua algoritma sebagai metode yang kuat dalam analisis data klinis.

Berdasarkan penelitian-penelitian tersebut, masih diperlukan kajian lebih lanjut yang secara spesifik membandingkan performa Regresi Logistik dan Random Forest pada dataset liver dengan fokus pada analisis faktor penting yang memengaruhi prediksi. Penelitian ini bertujuan untuk menganalisis model klasifikasi

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

dalam memprediksi penyakit liver serta mengevaluasi performa kedua algoritma menggunakan metrik akurasi, presisi, recall, dan F1-score. Selain itu, penelitian ini bertujuan mengidentifikasi variabel klinis yang paling berpengaruh dalam proses klasifikasi menggunakan model-machine learning, sehingga dapat menjadi dasar untuk sistem pendukung keputusan di bidang kesehatan.

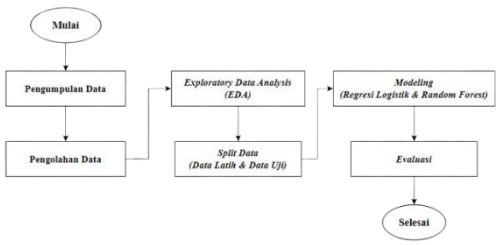
METODE PENELITIAN

Metode penelitian ini dirancang untuk menganalisis model klasifikasi dalam memprediksi penyakit liver menggunakan algoritma Regresi Logistik dan Random Forest. Penelitian diawali dengan proses pengumpulan data, yaitu dataset penyakit liver diperoleh dari platform Kaggle. Dataset ini berisi informasi pasien seperti usia, jenis kelamin, kadar bilirubin, albumin, dan enzim hati yang digunakan sebagai variabel prediktor. Setelah data terkumpul, dilakukan pengolahan data, yang meliputi pembersihan data (data cleaning) untuk mengatasi nilai hilang, penanganan duplikasi data, serta normalisasi atau standarisasi variabel numerik agar setiap fitur berada pada skala yang proporsional.

Tahap berikutnya adalah Exploratory Data Analysis (EDA) yang bertujuan untuk memahami karakteristik data secara lebih mendalam, seperti distribusi variabel, pola hubungan antar fitur, serta perbandingan jumlah pasien liver dan non-liver berdasarkan gender maupun rentang usia. Hasil EDA membantu dalam mengidentifikasi pola atau ketidakseimbangan data yang dapat memengaruhi performa model klasifikasi. Setelah itu, dilakukan proses pembagian data (split data) menjadi data latih dan data uji dengan proporsi 80:20. Pembagian ini dilakukan menggunakan teknik *stratified split* agar distribusi kelas pada data uji tetap seimbang dengan data asli sehingga proses evaluasi lebih representatif.

Selanjutnya dilakukan tahap pemodelan (modeling) menggunakan dua algoritma yaitu Regresi Logistik dan Random Forest. Regresi Logistik karena sifatnya yang interpretatif dan mampu menjelaskan pengaruh masing-masing variabel terhadap probabilitas penyakit liver. Sementara itu, Random Forest digunakan karena kemampuannya dalam menangani data berukuran besar, melakukan pemilihan fitur otomatis, serta menghasilkan performa prediksi yang stabil. Kedua model dilatih menggunakan data latih, dan hasil prediksinya dibandingan pada data uji.

Tahap penelitian tersebut divisualisasikan dalam bentuk diagram alir (flowchart) pada Gambar 1, yang menggambarkan urutan proses dari pengumpulan data hingga evaluasi model.



Gambar 1. Tahapan Penelitian

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

1. Pengumpulan Data

Pengumpulan data pada penelitian ini dilakukan dengan memanfaatkan dataset penyakit liver yang diperoleh dari platform Kaggle, yaitu repositori data publik yang banyak digunakan dalam penelitian data mining dan *machine learning*. Dataset ini terdiri dari 579 entri pasien dengan berbagai atribut klinis yang berkaitan dengan fungsi hati, di antaranya usia (age), jenis kelamin (gender), kadal Total Bilirubin, Proteins, Albumin, dan Albumin and Globulin Ratio. Selain itu, terdapat atribut dataset yang digunakan sebagai label kelas untuk membedakan pasien yang terindikasi penyakit liver (1) dan pasien yang tidak terindikasi penyakit liver (2).

Gambar dataset yang ditampilkan menunjukkan struktur data mentah yang menjadi dasar dalam proses analisis, di mana setiap baris merepresentasikan satu pasien dengan nilai klinis masing-masing. Dataset ini dipilih karena memiliki cakupan variabel klinis yang relevan serta formatnya yang terstruktur, sehingga sangat sesuai untuk dianalisis menggunakan teknik klasifikasi Regresi Logistik dan Random Forest dalam penelitian ini. Data kemudian diunduh dalam format CSV untuk diproses pada tahap selanjutnya.

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Dataset
0	65	0	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	1	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	1	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	1	1.0	0.4	182	14	20	6.8	3.4	1.00	1
4	72	1	3.9	2.0	195	27	59	7.3	2.4	0.40	1
5	46	1	1.8	0.7	208	19	14	7.6	4.4	1.30	1
6	26	0	0.9	0.2	154	16	12	7.0	3.5	1.00	1
7	29	0	0.9	0.3	202	14	11	6.7	3.6	1.10	1
8	17	1	0.9	0.3	202	22	19	7.4	4.1	1.20	2
9	55	1	0.7	0.2	290	53	58	6.8	3.4	1.00	1

Gambar 2. Tampilan Dataset Pasien Penyakit Liver

2. Pengolahan Data

Tahap pengolahan data dilakukan setelah dataset berhasil dikumpulkan untuk memastikan bahwa model berada dalam kondisi bersih, terstruktur, dan siap digunakan dalam proses analisis serta pemodelan. Langkah pertama pada tahap ini adalah melakukan pemeriksaan nilai hilang (missing values) dan duplikasi data, karena keberadaan data yang kosong atau rangkap dapat menurunkan kualitas model dan menghasilkan prediksi yang tidak akurat. Ketidaksempurnaan data kemudian ditangani dengan metode seperti pengisian menggunakan nilai median untuk variabel numerik atau penghapusan duplikasi untuk menjaga integritas dataset. Selanjutnya dilakukan proses transformasi variabel, misalnya mengubah variabel *Gender* dari bentuk numerik (0 dan 1) menjadi label kategorikal (*Perempuan* dan *Laki-laki*) agar lebih mudah dalam tahap eksplorasi data dan visualisasi.

Pengolahan data juga mencakup standarisasi fitur numerik, terutama kerena beberapa variabel seperti *Alkaline Phosphatase* dan *Aminotransferase* memiliki rentan nilai yang jauh lebih tinggi dibandingkan variabel lain seperti *Albumin* atau *Direct Bilirubin*. Standarisasi dilakukan menggunakan teknik *StandardScaler* agar setiap fitur memiliki skala yang setara sehingga algoritma seperti Regresi Logistik dapat bekerja secara optimal. Setelah itu, dilakukan pembuatan label kelas (target)

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

dengan memetakan nilai pada kolom Dataset menjadi dua kategori, yaitu *Liver* untuk pasien yang terindikasi penyakit liver dan *Non-Liver* untuk pasien yang tidak terindikasi. Tahap pengolahan data yang sistematis ini memastikan bahwa dataset layak untuk digunakan dalam proses *Exploratory Data Analysis* (EDA), pembagian data, dan pemodelan menggunakan algoritma Regresi Logistik dan Random Forest. Dengan demikian, kualitas data yang terjaga memberikan kontribusi penting terhadap akurasi dan stabilitas hasil penelitian.

```
# Cek missing value
print("\nCek missing value:")
print(df.isnull().sum())

# Jika ada missing value → isi dengan median
df = df.fillna(df.median(numeric_only=True))

# Cek duplikasi
print("\nJumlah data duplikat:", df.duplicated().sum())

# Hapus jika ada duplikat
df = df.drop_duplicates()
```

Gambar 3. Pembersihan Data

```
Cek missing value:
Age                             0
Gender                          0
Total_Bilirubin                 0
Direct_Bilirubin                0
Alkaline_Phosphotase            0
Alamine_Aminotransferase        0
Aspartate_Aminotransferase      0
Total_Protiens                  0
Albumin                         0
Albumin_and_Globulin_Ratio      0
Dataset                         0
dtype: int64

Jumlah data duplikat: 13
```

Gambar 4. Hasil Pembersihan Data

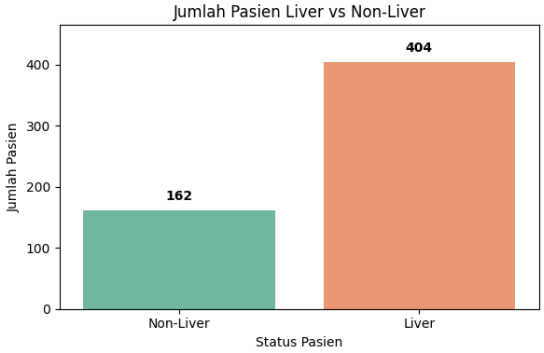
Exploratory Data Analysis (EDA)

Tahap *Exploratory Data Analysis* (EDA) dilakukan untuk memahami karakteristik data, melihat pola distribusi, serta mengidentifikasi perbedaan antara pasien yang terindikasi penyakit liver dan yang tidak. Analisis eksporatif ini sangat penting karena memberikan gambaran awal mengenai kondisi dataset sebelum dilakukan pemodelan lebih lanjut. Pada tahap ini, beberapa visualisasi data digunakan untuk memperjelas pola dan kecenderungan yang muncul dari variabel-variabel yang dianalisis

Dalam penelitian ini, terdapat 3 (tiga) jenis analisis yang dilakukan pada tahap EDA, yaitu:

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

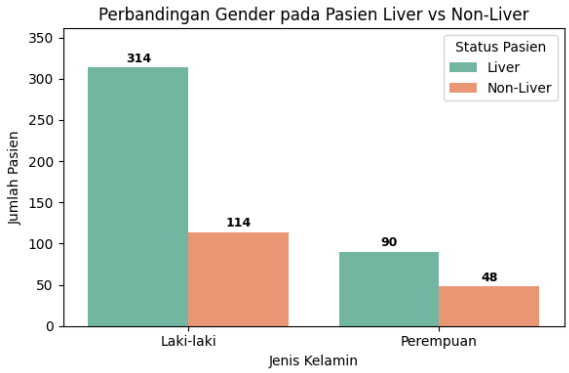
1. EDA Status Pasien Liver



Gambar 5. Jumlah Pasien Liver vs Non-Liver

Analisis pertama dilakukan dengan melihat distribusi jumlah pasien yang terindikasi penyakit liver dan yang tidak terindikasi. Berdasarkan visualisasi pada Gambar 5, terlihat bahwa dataset memiliki ketidakseimbangan kelas (*class imbalance*) yang cukup signifikan. Jumlah pasien yang terdiagnosis liver mencapai 404 orang, sedangkan pasien non-liver hanya berjumlah 162 orang. Perbedaan ini menunjukkan bahwa kelas *Liver* menjadi kelas mayoritas, sementara kelas *Non-Liver* berada pada proporsi yang jauh lebih kecil. Kondisi seperti ini penting diperhatikan karena dapat memengaruhi kinerja model klasifikasi, terutama pada algoritma yang sensitif terhadap ketidakseimbangan data. Model cenderung lebih mudah memprediksi kelas mayoritas dan berpotensi mengabaikan kelas minoritas. Oleh karena itu, hasil analisis ini menjadi dasar bagi peneliti untuk lebih cermat dalam mengevaluasi model menggunakan metrik yang tidak hanya melihat akurasi, tetapi juga precision, recall, dan F1-Score agar penilaian performa lebih objektif.

2. EDA Berdasarkan Jenis Kelamin



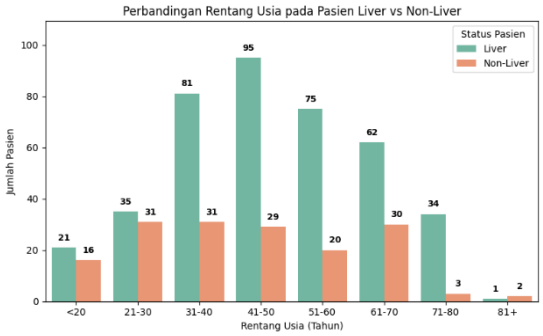
Gambar 6. Perbandingan Gender pada Pasien Liver vs Non-Liver

Analisis selanjutnya dilakukan dengan membandingkan distribusi pasien liver dan non-liver berdasarkan jenis kelamin. Visualisasi pada gambar 6, menunjukkan bahwa adanya perbedaan yang

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

cukup mencolok antara laki-laki dan perempuan. Pasien laki-laki tercatat sebagai kelompok yang paling banyak mengalami penyakit liver dengan jumlah 314 pasien, sedangkan perempuan hanya berjumlah 90 pasien. Pada kategori non-liver, laki-laki juga mendominasi dengan 114 pasien, sementara perempuan berjumlah 48 pasien. Temuan ini menunjukkan bahwa laki-laki memiliki kecenderungan yang lebih tinggi untuk mengalami gangguan fungsi hati dibandingkan perempuan. Faktor biologis, gaya hidup, pola konsumsi alkohol, serta kebiasaan merokok yang lebih sering ditemukan pada laki-laki dapat menjadi penyebab dominan kasus liver pada kelompok ini, sebagaimana juga dilaporkan dalam beberapa penelitian medis. Dengan demikian, variabel gender terbukti menjadi salah satu fitur yang relevan dalam proses klasifikasi penyakit liver.

3. EDA Rentan Usia



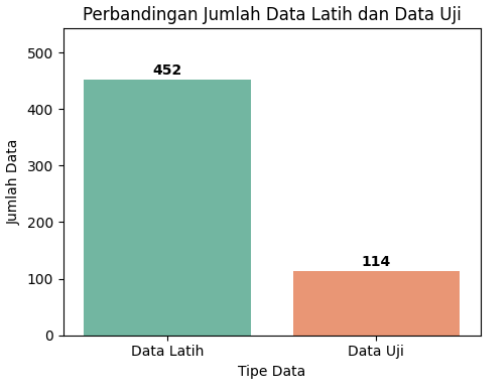
Gambar 7. Perbandingan Rentan Usia pada Pasien Liver vs Nonliver

Analisis berikutnya dilakukan untuk mengetahui bagaimana distribusi penyakit liver berdasarkan rentan usia. Data kemudian dikelompokkan menjadi beberapa kategori usia, mulai dari usia kurang dari 20 tahun hingga lebih dari 80 tahun. Berdasarkan visualisasi pada Gambar 7, kelompok usia 41-50 tahun menjadi kategori dengan jumlah kasus tertinggi, yaitu 95 pasien. Disusul oleh kelompok usia 31-41 tahun dengan 81 pasien, kemudian usia 51-60 tahun sebanyak 75 pasien. Sementara itu, kelompok usia di bawah 20 tahun dan di atas 80 tahun memiliki jumlah kasus yang sangat rendah sehingga tidak menunjukkan kontribusi signifikan dalam pola penyakit liver. Pola ini mengindikasikan bahwa risiko penyakit liver cenderung meningkat pada usia dewasa pertengahan hingga lanjut, di mana fungsi organ mulai mengalami penurunan dan gaya hidup jangka panjang mulai memberikan dampak. Dengan demikian, rentan usia juga menjadi faktor penting yang perlu diperhatikan dalam proses klasifikasi, karena memberikan informasi risiko yang berbeda pada setiap kelompok usia.

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

3. Split Data

Tahap split data dilakukan untuk membagi dataset menjadi dua bagian utama, yaitu data latih (training data) dan data uji (testing data). Pembagian ini bertujuan untuk memastikan bahwa model dapat dilatih menggunakan sebagian data, kemudian dievaluasi menggunakan data yang belum pernah dilihat sebelumnya, sehingga performanya lebih objektif dan tidak mengalami overfitting.



Gambar 8. Perbandingan Jumlah Data Latih dan Data Uji

Pada penelitian ini, data dibagi dengan proporsi 80% untuk data latih dan 20% untuk data uji, sebagaimana ditunjukkan pada Gambar 8. Dari total 556 data yang tersisa setelah proses pembersihan, sebanyak 452 data digunakan sebagai data latih dan 114 data digunakan sebagai data uji. Selain itu, proses pembagian data dilakukan menggunakan teknik stratified split, yaitu metode pembagian data yang mempertahankan proporsi kelas *Liver* dan *Non-Liver* pada kedua bagian dataset. Teknik ini penting untuk menghindari ketidakseimbangan distribusi kelas yang dapat menyebabkan model bias terhadap kelas tertentu. Dengan demikian, proses split data memastikan bahwa model dapat dilatih secara optimal dan diuji secara adil, sehingga hasil evaluasi yang diperoleh lebih valid dan representatif.

4. Modeling

Tahap *modeling* dilakukan untuk membangun model klasifikasi yang mampu memprediksi apakah seseorang pasien terindikasi penyakit liver atau tidak berdasarkan variabel klinis yang tersedia. Pada penelitian ini digunakan dua algoritma yaitu Regresi Logistik dan Random Forest, yang masing-masing memiliki karakteristik berbeda namun sama-sama efektif dalam menyelesaikan masalah klasifikasi biner. Sebelum model dibangun, seluruh fitur numerik telah distandarisasi menggunakan StandardScaler agar variabel memiliki skala yang beragam, sehingga proses pelatihan model menjadi lebih stabil dan akurat.

Algoritma Regresi Logistik digunakan sebagai model pertama karena sifatnya yang sederhana, interpretatif, dan mampu menunjukkan kontribusi masing-masing variabel terhadap probabilitas terjadi penyakit liver. Model ini bekerja dengan menghitung hubungan linier antara variabel prediktor dan probabilitas klasifikasi, kemudian mengubah nilai antara 0 dan 1 melalui fungsi sigmoid. Kelebihan utama dari Regresi Logistik adalah kemudahan interpretasi serta kinerjanya yang baik pada

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

dataset terstandarisasi, sehingga sangat sesuai digunakan sebagai model dasar (baseline model) dalam analisis klasifikasi.

```
# Re-run Logistic Regression model training
log_model = LogisticRegression(max_iter=1000)
log_model.fit(X_train, y_train)
y_pred_log = log_model.predict(X_test)
y_pred_log_prob = log_model.predict_proba(X_test)[:, 1]
```

Gambar 9. Pemodelan Regresi Logistik

Model kedua yang digunakan adalah Random Forest, yaitu algoritma berbasis *ensemble learning* yang membangun sejumlah pohon keputusan (decision trees) dan menggabungkan hasil prediksinya untuk menghasilkan keputusan akhir yang lebih akurat dan stabil. Random Forest memiliki kemampuan untuk menangani variabel dalam jumlah banyak, bersifat tahan terhadap *overfitting*, serta mampu melakukan pemilihan fitur secara otomatis melalui mekanisme *feature importance*. Model ini sangat efektif pada dataset dengan keragaman nilai yang tinggi, seperti variabel enzim hati, kadar protein, dan parameter klinis lainnya pada dataset liver.

```
# Re-run Random Forest model training
rf_model = RandomForestClassifier(n_estimators=200, random_state=42)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
y_pred_rf_prob = rf_model.predict_proba(X_test)[:, 1]
```

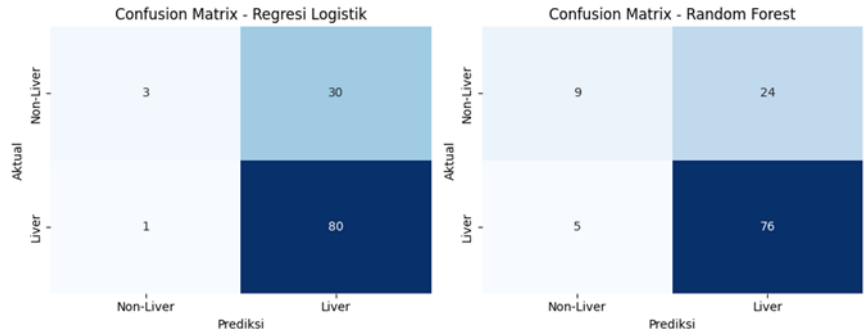
Gambar 10. Pemodelan Random Forest

Kedua model dilatih menggunakan data latih (training set) dan menghasilkan prediksi terhadap data uji (testing set). Hasil prediksi dari kedua model ini kemudian dibandingkan menggunakan beberapa metrik evaluasi untuk mengetahui model mana yang memiliki performa terbaik dalam mengklasifikasi penyakit liver. Dengan penggunaan dua pendekatan yang berbeda ini, penelitian dapat melihat perbandingan performa antara model yang bersifat linear dan model berbasis *ensemble*, sehingga kesimpulan yang diperoleh lebih komprehensif.

HASIL DAN PEMBAHASAN

Evaluasi model dilakukan untuk menilai performa algoritma Regresi Logistik dan Random Forest dalam mengklasifikasikan pasien liver dan non-liver. Penilaian dilakukan menggunakan confusion matrix dan beberapa metrik evaluasi, yaitu accuracy, precision, recall, dan F1-score. Confusion matrix memberikan gambaran mengenai distribusi hasil prediksi model terhadap data uji, termasuk prediksi benar maupun kesalahan model dalam mengidentifikasi kelas Liver dan Non-Liver.

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930



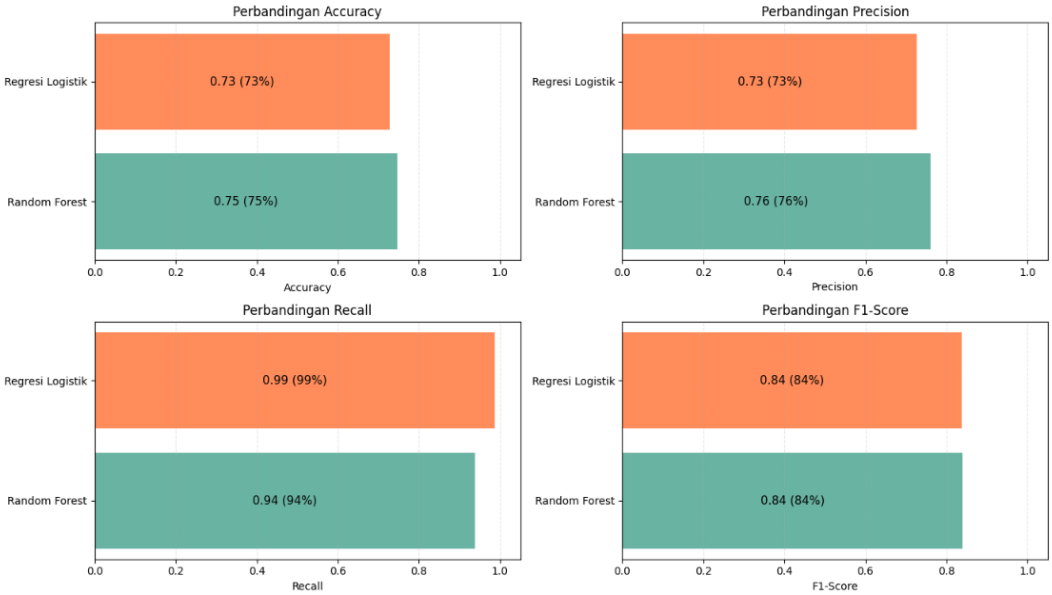
Gambar 11. Confusion Matrix Regresi Logistik dan Random Forest

Confusion matrix model Regresi Logistik, terlihat bahwa model ini sangat baik dalam mendeteksi pasien yang benar-benar memiliki penyakit liver. Hal tersebut ditunjukkan oleh jumlah True Positive (80) yang cukup tinggi dan False Negative (1) yang sangat rendah, sehingga hanya sedikit kasus liver yang terlewat oleh model. Namun, model ini memiliki kelemahan besar pada kemampuan mengenali pasien Non-Liver, karena jumlah True Negative hanya 3, sedangkan False Positive mencapai 30. Kondisi ini menunjukkan bahwa model cenderung memberikan prediksi "Liver" meskipun pasien sebenarnya tidak memiliki penyakit liver, sehingga terjadi banyak false alarm. Dengan demikian, Regresi Logistik tampak bias ke kelas Liver dan tidak seimbang dalam membedakan kedua kelas.

Sementara itu, model Random Forest menunjukkan performa yang lebih seimbang antara mendeteksi pasien Liver dan Non-Liver. Meskipun jumlah True Positive sedikit lebih rendah daripada regresi logistik (76 dibandingkan 80), model ini memiliki True Negative yang lebih tinggi, yaitu 9, serta False Positive yang lebih rendah, yaitu 24. Hal ini berarti Random Forest lebih baik dibandingkan regresi logistik dalam mengenali kasus Non-Liver. Akan tetapi, Random Forest menghasilkan False Negative sebanyak 5, yang berarti ada lebih banyak pasien liver yang tidak terdeteksi dibanding regresi logistik. Meskipun begitu, model ini tetap lebih stabil dan tidak terlalu bias terhadap salah satu kelas. Secara keseluruhan, regresi logistik unggul dalam meminimalkan kesalahan pada kasus liver, sedangkan Random Forest lebih seimbang karena mampu menurunkan jumlah false alarm dan meningkatkan deteksi Non-Liver.

VISA: Journal of Visions and Ideas
Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

Selain visualisasi confusion matrix, perbandingan matrix evaluasi dari kedua model ditunjukkan pada Gambar 12.



Gambar 12. Perbandingan Accuracy, Precision, Recall, F1-Score

Selain confusion matrix, kinerja kedua model dievaluasi menggunakan metrik accuracy, precision, recall, dan F1-Score untuk memberikan gambaran performa yang lebih komprehensif. Berdasarkan hasil pengujian, model Regresi Logistik memperoleh nilai akurasi 0,73, yang menunjukkan kemampuan model dalam memberikan prediksi benar sebesar 73%. Model ini juga memiliki nilai recall sangat tinggi sebesar 0,99, yang berarti hampir seluruh pasien liver berhasil teridentifikasi. Namun, nilai precision yang lebih rendah menunjukkan bahwa ada banyak pasien non-liver yang salah diklasifikasi sebagai liver. Hal ini sejalan dengan tingginya nilai *false positive* pada confusion matrix.

Di sisi lain, model Random Forest menunjukkan performa yang lebih seimbang, dengan nilai akurasi 0,75, sedikit lebih tinggi dibandingkan Regresi Logistik. Nilai precision 0,76 yang lebih baik menggambarkan bahwa model lebih tepat dalam mengidentifikasi pasien liver tanpa banyak menghasilkan kesalahan prediksi pada pasien non-liver. Namun, nilai recall Random Forest sedikit lebih rendah dibandingkan Regresi Logistik, menunjukkan bahwa model ini sedetail Regresi Logistik dalam menangkap seluruh kasus liver. Meskipun demikian, kedua model memiliki F1-Score yang sama yaitu 0,84, menandakan keseimbangan yang baik antara recall dan precision.

Secara keseluruhan, perbandingan metrik evaluasi menunjukkan bahwa Random Forest memberikan performa yang lebih konsisten, terutama dalam mengurangi kesalahan prediksi terhadap kelas non-liver. Sementara itu, Regresi Logistik sangat unggul dalam mendeteksi kasus liver sehingga cocok digunakan pada skenario deteksi dini yang mengutamakan sensitivitas.

Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang dilakukan mengenai analisis model klasifikasi untuk prediksi penyakit liver menggunakan algoritma Regresi Logistik dan Random Forest, dapat disimpulkan bahwa kedua model memiliki kinerja yang baik dalam mengidentifikasi pasien liver dan non-liver. Model Regresi Logistik menunjukkan kemampuan deteksi yang sangat tinggi terhadap kasus positif, terlihat dari nilai recall sebesar 0,99, sehingga efektif digunakan untuk keperluan deteksi dini. Namun demikian, model ini memiliki keterbatasan dalam ketepatan prediksi terhadap kelas non-liver akibat tingginya nilai false positive. Sementara itu, algoritma Random Forest memberikan performa yang lebih seimbang dan stabil dengan nilai akurasi 0,75 serta precision 0,76, sehingga lebih handal dalam mengurangi kesalahan prediksi pada kelas negatif. Kedua model menghasilkan nilai F1-Score yang sama yaitu 0,84, menunjukkan adanya keseimbangan antara precision dan recall. Dengan demikian, pemilihan model perlu disesuaikan dengan kebutuhan implementasi, apakah lebih mengutamakan sensitivitas (Regresi Logistik) atau Stabilitas Prediksi (Random Forest).

Untuk pengembangan penelitian di masa mendatang, disarankan agar cakupan data diperluar guna meningkatkan kemampuan generalisasi model. Peneliti juga dianjurkan untuk menambahkan algoritma pembanding lainnya, seperti Support Vector Machine (SVM), Gradient Boosting, atau XGBoost, sehingga diperoleh model dengan performa yang lebih optimal. Penggunaan teknik *hyperparameter tuning* juga penting dilakukan untuk meningkatkan kualitas model secara signifikan. Selain itu, hasil model prediksi ini berpotensi dikembangkan menjadi sistem pendukung keputusan medis berbasis aplikasi, sehingga dapat membantu tenaga kesehatan dalam proses diagnosis awal penyakit liver. Terakhir, penelitian lanjutan dapat memanfaatkan pendekatan analisis fitur seperti *feature importance* atau *SHAP analysis* untuk memahami variabel-variabel yang memberikan pengaruh terbesar terhadap prediksi, sehingga hasil penelitian dapat memberikan kontribusi yang lebih komprehensif pada bidang medis dan kesehatan masyarakat.

DAFTAR PUSTAKA

An, M. E., Griffin, P., Stine, J. G., State Health, P., Balakrishnan, R., & Kumara, S. (2025). *Predicting Metabolic Dysfunction-Associated Steatotic Liver Disease using Machine Learning Methods*.

Herliawan, I., Muhammad Iqbal, ;, Gata, ; Windu, Rifai, A., Jajang, ;, Purnama, J., Science, C., Mandiri, S. N., & Id, W. N. A. (2020). *CLASSIFICATION OF LIVER DISEASE BY APPLYING RANDOM FOREST ALGORITHM AND BACKWARD ELIMINATION*. *6*(1). https://doi.org/10.33480/jitk.v6i1.1424

Hikmah, A. B. (2025). Optimasi Hyperparameter Ensemble Learning untuk Prediksi Penyakit Liver Berdasarkan Data Pasien. *Jurnal Sistem Komputer dan Kecerdasan Buatan, VIII*.

Karna, A., Khan, N., Rauniyar, R., & Shambharkar, P. G. (2024). *Unified dimensionality reduction techniques in chronic liver disease detection*. http://arxiv.org/abs/2412.21156

Vol 6 No 1 (2026) 418–430 E-ISSN 2809-2058 P-ISSN 2809-2643
DOI: 47467/visa.v6i1.10930

Lu, J. (2023). Research on Prediction of Liver Disease Based on Machine Learning Models. In *Highlights in Science, Engineering and Technology ISET* (Vol. 2023).

Mardewi. (2023). *PREDIKSI DINI LIVER CIRRHOSIS UNTUK KESEHATAN HATI MENGGUNAKAN METODE MACHINE LEARNING*. https://www.kaggle.com/datasets/fatemehmehrparvar/liver-

Nurkholifah, M., Jasmarizal, Umar, Y., & Rahmaddeni. (2023). ANALISA PERFORMA ALGORITMA MACHINE LEARNING DALAM PREDIKSI PENYAKIT LIVER. *Jurnal Indonesia : Manajemen Informatika Dan Komunikasi*, *4*(1), 164–172. https://doi.org/10.35870/jimik.v4i1.149

Sen, S. K., Kumar, S., Swain, N. K., & Mitra, S. P. (2025). Automated Liver Disease Diagnosis using Machine Learning Techniques. In *Journal of Neonatal Surgery ISSN* (Vol. 14, Issue 2). Shankar Prasad Mitra. https://www.jneonatalsurg.com

Silmi Ath Thahirah Al Azhima1, D. D. N. F. A. H. I. K. M. A. Q. N. S. S. (2022). arahmah,+49_6_HYBRID+MACHINE+LEARNING+MODEL+UNTUK+MEMPREDIKSI+PENYAKIT+JANTUNG+PADA+SISTEM+INFORMASI+BERBASIS+DEKSTOP. *Jurnal Teknologi Terpadu*.

Zhang, H., Zhang, L., Li, N., Zhang, Y., Zhang, X., & Wang, D. (2025). Machine learning survival models for Non-alcoholic fatty liver disease based on a health checkup cohort. *BMC Gastroenterology*, *25*(1). https://doi.org/10.1186/s12876-025-04120-6