KLASIFIKASI DATA MINING DENGAN ALGORITMA MACHINE LARNING UNTUK PREDIKSI PENYAKIT LIVER

F. Lia Dwi Cahyanti[1)], Fajar Sarasati[2)], Widi Astuti [3)], Elly Firasari [4)]

[1]Fakultas Teknologi Informasi, Universitas Nusa Mandiri
email: [1]flia.fdc@nusamandiri.ac.id
[2] Fakultas Ekonomi dan Bisnis, Universitas Nusa Mandiri
email: [2]fajar.fss@nusamandiri.ac.id
[3] Fakultas Ekonomi dan Bisnis, Universitas Nusa Mandiri
email: [3]widiastuti.wtu@nusamandiri.ac.id
[4] Fakultas Teknologi Informasi, Universitas Nusa Mandiri
email: [4]elly.efa@nusamandiri.ac.id

Abstrak

Liver merupakan organ tubuh manusia yang memiliki peranan sangat penting seperti mencerna, menyerap, membantu proses pencernaan makanan serta menghancurkan racun di dalam darah. Penyakit hati atau liver yang sudah akut sangat mempengaruhi fungsi-fungsi hati, penyakit hati dapat diketahui dari munculnya gejala klinis maupun fisik yang timbul pada pasien. Penelitian ini membahas tentang klasifikasi penyakit liver pada dataset ILPD yang diambil dari UCI *Machine learning Repository* menggunakan algoritma *machine learning.* Dataset terdiri dari 583 record data, 10 kriteria, dan 1 variable kelas berjenis multivariate. Penelitian ini menggunakan beberapa tahapan preprocessing yang dilakukan, diantaranya : Preprocessing Data Dan Eksplorasi Data, Penanganan *missing value, feature selection*, menerapkan feature correlation dan feature scaling, Analisis menggunakan Algoritma *Machine learning.* Berdasarkan hasil pengujian yang dilakukan dalam memperoleh nilai akurasi perhitungan klasifikasi menggunakan Algoritma Random Forest memiliki performa keakuratan yang diukur dengan akurasi sebesar 78,63% sehingga disimpulkan akurasi tersebut lebih unggul dari algoritma lainnya dalam klasifikasi penyakit liver.

Kata kunci: Klasifikasi*, Machine Learning, Random Forest, Liver.*

1. PENDAHULUAN

Liver merupakan organ tubuh manusia yang sangat penting yang terletak di sebelah kanan sisi perut, mempunyai berat sekitar 3 pon dan berwarna coklat kemerahan. Liver memiliki dua bagian besar yang disebut lobus kanan dan kiri. Liver memiiki fungsi untuk mencerna, menyerap, membantu proses pencernaan makanan serta menghancurkan racun di dalam darahdengan cara mendetoksifikasi bahan kimia untuk mengeluarkan racun dan fungsi lainnya (Falatehan, Hidayat dan Brata, 2018).

Penyakit hati yang sudah akut sangat mempengaruhi fungsi-fungsi hati, penyakit hati dapat diketahui dari munculnya gejala klinis maupun fisik yang timbul pada pasien, Gejala klinis dapat diketahui dari apa yang dirasakan oleh pasien, sedangkan gejala fisik dapat diketahui dari keadaan tubuh pasien, Gejala penyakit hati ada banyak dan kompleks, serta penyakit hati memiliki kemiripan gejala dengan beberapa penyakit (Falatehan, Hidayat dan Brata, 2018).

Menurut WHO (World Health Organization) tahun 2013, liver merupakan penyakit yang dianggap sebagai pembunuh diam-diam tanpa gejala. pasien penderita penyakit liver di Indonesia mencapai 28 juta orang hal tesebut membuat penyakit liver disebut sebagai salah satu dari 10 penyakit dengan tingkat kematian yang paling tinggi sehingga angka kematian setiap tahun semakin meningkat (Widodo, Rohman dan Sismindari, 2019).

Kendala dan masalah yang terjadi pada seserorang yaitu sulitnya mengenali gejala penyakit liver sejak dini. Mengetahui adanya gejala penyakit liver sejak dini sangatlah penting agar pasien penderita penyakit liver dapat lebih awal meningkatkan tingkat kelangsungan hidupnya. Maka dari itu penelitian ini difokuskan untuk mengetahui model manakanh yang paling sesuai untuk mengklasifikasikan data penyakit liver , sehingga penyakittersebut bisa dideteksi lebih awal (Lin, 2009).

Pada bidang ilmu komputer, data mining merupakan metode yang dapat digunakan untuk menganalisis kumpulan data. Algoritma *machine learning* adalah suatu metode pembelajaran mesin yang mengacu pada teknik yang berhubungan dengan pola berdasarkan model untuk kalsifikasi dan prediksi data baru. Pada prinsipnya, *machine learning* memiliki empat langkah : definisi masalah, pengumpulan data dan persiapan data, pembuatan model dan prediksi model (Zaidan *et al.*, 2020). algoritma *machine learning* dapat digunakan untuk memecahkan masalah sesuai dengan kebutuhan pada bidang masing-masing (Roihan, Sunarya dan Rafika, 2020).

Beberapa peneletian terdahulu yang berkaitan dengan penelitian ini yaitu seperti penelitian yang dilakukan oleh rahman th 2020 melakukan penelitian mengenai penyakit liver menggunakan metode decision tree dan naive bayes yang mempunyai nilai optimal akurasi sebesar 70.29% (Rahman, 2020). Lubis, dkk juga melakukan penelitian sebelumnya dengan random forest dan naive bayes dalam memprediksi penyakit liver (Lubis, Erdiansyah dan Siregar, 2022). Berdasarkan penelitian yang telah dilakukan peneliti sebelumnya dengan metode Random Forest memperoleh hasil yaitu akurasi sebesar 70.60 % yang didapatkan hanya memiliki tapan preprocesing normalisasi data saja sehingga ini mencerminkan bahwa model klasifikasi yang didapat tidakalah bagus. Pada penelitian ini peneliti akan membandingkan klasifikasi penyakit liver dengan menggunakan tahapan preprocessing yang lebih kompleks diantaranya pengecekan missing value, imputasi, feature selection, dan resampling untuk mengatasi imbalance data. Oleh karena itu, pada penelitian ini metode yang akan digunakan peneliti untuk melakukan klasifikasi pada Indian liver dataset adalah algoritma *machine learning*.

2. LANDASAN TEORI

A. *Machine learning*

Machine learning (Pembelajaran Mesin) adalah suatu mesin yang dikembangkan dengan kemampuan komputer untuk melakukan pembelajaran sendiri tanpa ada arahan [9]. Pembelajaran mesin (ML) adalah studi ilmiah mengenai algoritma dan model statistik sistem komputer yang digunakan untuk melakukan tugas tertentu tanpa diprogram secara eksplisit (Friedman, 2001).

B. Data Mining

Data mining adalah proses mencari informasi dengan cara mengidentifikasi pola dari data penting dan memprediksi output, Data mining juga disebut pengetahuan penemuan dalam basis data yang diimplementasikan menggunakan berbagai jenis data seperti database relational, gudang data, repositori Data dan sebagainya. Data mining terbukti menjadi pendekatan yang kuat dan efektif yang menyediakan proses penemuan pola dalam dataset besar (Prajarini, 2016).

C. Klasifikasi

Klasifikasi adalah proses untuk menemukan sebuah model berdasarkan kelas-kelas yang digunakan sebagai pembeda antara kelas satu dengan kelas yang lain. Klasifikasi melakukan pembangunan model berdasarkan data latih yang ada, kemudian menggunakan model tersebut untuk mengklasifikasikan pada data yang baru. Klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/ pembelajaran terhadap fungsi target yang memetakan setiap set atribut (fitur) ke satu jumlah label kelas yang tersedia (Utomo, 2020).

D. Random Forest

Menurut (Azis, Tangguh Admojo dan Susanti, 2020) Random forests adalah kombinasi 3 pohon keputusan yang mana setiap pohon bertanggung jawab pada setipa nilai dari random vector yang diberikan begitupun persebarannya akan diberikan kepada pohon keputusan lainnya.

Random forest merupakan salah satu model klasifikasi yang menerapkan pohon keputusan tanpa memotong pohon untuk memaksimalkan algoritma dan akurasi yang didapat untuk menghindari overfitting. Proses random forest dilakukan dengan melakukan rancangan sampling bootstrap sampling dengan melakukan Replace (Wu *et al.*, 2017). Gambaran model random forest dapat dilihat pada Gambar 1 sebagai berikut.

Algorithm 1 Ensemble Random Forest Algorith

Input: Training data T, parameters $\{\lambda, \delta, k, s,$
Output: Model with evaluation
1. Ensemble-RF$(T, \lambda, \delta, k, s, c)$
2. for $i < -1$ *to* s **do**
3. $(train,test) \leftarrow$ randomSplit(T, λ)
4. $split \leftarrow$ bootstrap$(train, \delta, k)$
5. $model \leftarrow RandomForest$.train$(split, c)$
6. $score \leftarrow$ evaluate$(model, test)$
7. out$[i] \leftarrow (model, score)$
8. end for
9. return *out*

Sumber : Weiwei Lin, dkk
Gambar 1. Proses Model Random Forest

3. METODE PENELITIAN

Data yang dipakai dalam penelitian ini yaitu dataset ILDP (Indian Liver Patient Dataset) yang diambil dari halaman website UCI *Machine learning* Repository yang merupakan data sekunder dari website Uci *Machine learning* dengan link https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset). dengan data sebanyak 583 record dan 10 atribute.

Tabel 1. Data Variabel Penelitian

No	Variabel	Keterangan	Jenis Data
1	Age	Umur	Numerik
2	Gender	Jenis Kelamin	Kategori
3	TB	Total Bilirubin	Numerik
4	DB	Direct Bilirubin	Numerik
5	Alkaline	Alkphos Alkaline Phosphotase	Numerik
6	Alamine	Sgpt Alamine Aminotransferase	Numerik
7	Aspartate	Sgot Aspartate Aminotransferase	Numerik
8	TP	Total Protiens	Numerik
9	ALB	Albumin	Numerik
10	A/G	Rasio Albumin and Globulin	Numerik
11	Class	Menderita Liver/Tidak Menderita Liver	Kategori

Dataset tersebut akan dianalisis denganbeberapa langkah diantaranya :

1. Melakukan tahap preprocessing data seperti pengecekan *missing value, imputasi, feature selection,* dan *resampling* untuk mengatasi data *imbalance* data
2. Membagi dataset kedalam data *training* sebesar 80% dan *testing* sebesar 20% .
3. Melakukan klasifikasi dengan algoritma *machine learning* menggunakan tools python.
4. Memilih metode terbaik dengan hasil terbaik.

4. HASIL DAN PEMBAHASAN

A. Preprocessing Data Dan Eksplorasi Data

Pada tahap ini preprocessing data akan dilakukan pada seluruh variabel dalam dataset. Tahap pertama yang akan dilakukan yaitu pengecekan missing value dimana data apakah mempunyai nilai yang hilang atau tidak. Setelah dilakukan pengecekan ditemukan variabel yang mengalami kasus missing value yaitu 4 observasi pada variabel Albumin_and_Globulin_Ratio (A/G).

```
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  4
Dataset                     0
dtype: int64
```

Gambar 2. Atribut Missing Value

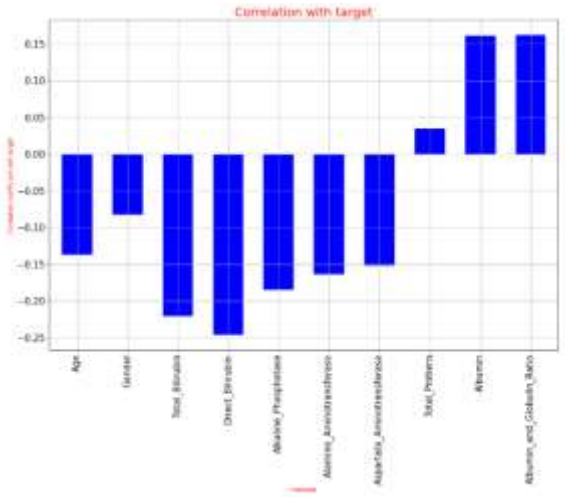
Permasalahan missing value akan diatasi dengan cara mengimputasikan nilai media kedalam variabel tersebut. Nilai media dinilai lebih robust dibandingan dengan nilai mean maupun modus. Variabel yang memiliki nilai yang hilang merupakan variabel yang mempunyai skala numerik.

```
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
```

Gambar 3. Atribut Setelah Proses Imputasi Nilai

Pada Gambar 3 dapat dilihat bahwa fitur Albumin_and_Globulin_Ratio yang sebelumnya memiliki nilai missing value sebanyak 4 sudah berubah menjadi 0 atau bisa dikatakan nilai missing value sudah terisi dengan imputasi nilai median.

Tahapan selanjutnya yaitu feature selection yang berfungsi untuk mengetahui fitur atau variabel apa saja yang mempengaruhi seseorang menderita penyakit liver. Nilai corelasi dengan target dapat dilihat pada gambar sebagai berikut :



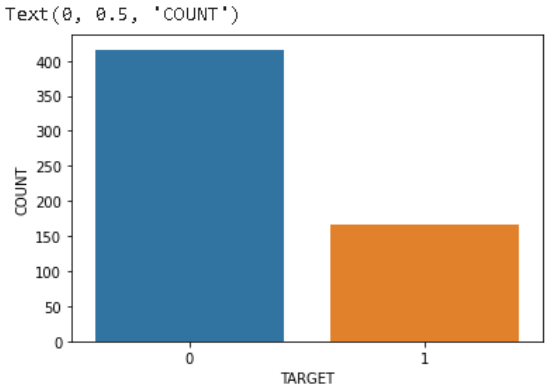
Gambar 4. Variabel Paling Berpengaruh

Pada Gambar 4. dapat dilihat bahwa nilai variabel important untuk gender (jenis kelamin) memiliki nilai mendekati class 0 , yang artinya gender tidak mempunyai pengaruh yang besar terhadap kalsifikasi penentuan apakah seseorang menderita penyakit liver atau tidak menderita penyakit liver, oleh sebab itu variabel gender akan dihapus dan tidak diikutkan dalam analisis selanjutnya. Selanjutnya akan dilakukan tahapan eksplorasi data yaitu dengan melihat hasil statistik pada variabel yang lainnya. Dilihat pada tabel 2 sebagai berikut :

Tabel 2. Statistika Variabel

Variabel	Mean	Std	Min	Max
Age	44.746141	16.189833	4.000000	90.000000
TB	3.298799	6.209522	0.400000	75.000000
DB	1.486106	2.808498	0.100000	19.700000
Alkaline	290.576329	242.937989	63.000000	2110.000000
Alamine	80.713551	182.620356	10.000000	2000.000000
Aspartate	109.910806	288.918529	10.000000	4929.000000
TP	6.483190	1.085451	2.700000	9.600000
ALB	3.141852	0.795519	0.900000	5.500000
A/G	0.946947	0.318495	0.300000	2.800000

Proses selanjutnya dilakukan eksplorasi variabel target pada penyakit liver, data dapat dilihat pada Gambar 5.

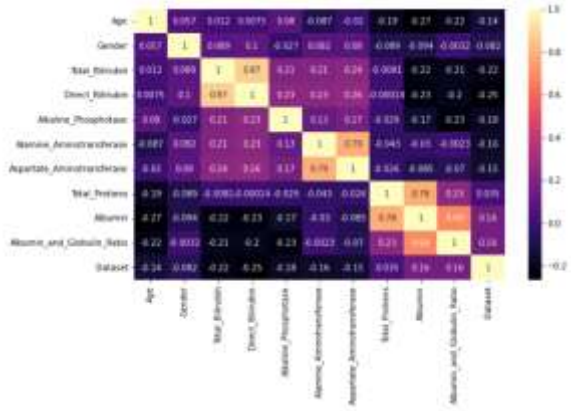


Gambar 5. Jumlah variabel target

Imbalance data pada gambar 5. Menunjukkan bahwa data dengan nilai 0 merupakan orang yang menderita penyakit liver sedangkan data dengan nilai 1 merupakan orang yang tidak menderita penyakit liver.

B. *Feature Correlation*

Feature Correlation digunakan untuk melihat korelasi antar atribut terhadap. Koefisien dari korelasi memiliki rentang nilai -1 hingga 1. Jika hasil korelasi mendekati angka 0, maka atribut tersebut memiliki korelasi yang lemah, Jika hasil korelasi mendekati angka -1, maka atribut memiliki korelasi negatif yang besar. Sedangkan korelasi yang mendekati dengan angka 1, maka atribut tersebut memiliki korelasi positif. Feature Correlation ditunjukkan pada gambar 6 sebagai berikut.



Gambar 6. *Feature Correlation*

Feature Correlation pada Gambar 6 dilihat berdasarkan fitur atau atribut yang berpengaruh dengan class. Atribut yang memiliki korelasi positif yaitu total_protiens, albumin, albumin_and_Globulin_ratio. Atribut yang memiliki korelasi negatif yaitu Age, Gender, Total_Bilitubin, Direct_Billirubin, Alkaline_Phosphatase, Alamine_Aminotransferase, Aspartate_Aminotransferase.

C. *Feature scaling*

Feature scaling terhadap dimensi data yang memiliki rentang nilai (scale) dengan perbedaan yang lebih menonjol dari nilai fitur yang lainnya, dapat dilihat bahwa atribut alkaline_Phosphotase, Alamine_Aminotransferase, Aspartate_Aminotransferase memiliki rentang nilai yang lebih menonjol atau bisa dikatakan lebih besar dari atribut lainnya.

Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase
187	16	18
699	64	100
490	60	68
182	14	20
195	27	59

Gambar 7. Feature yang di lakukan Scaling

Sehingga dilakukan scaling untuk membuta nilai lebih seimbang dari atribut yang lain. Berikut nilai fitur yang sudah melalui proses scaling, sehingga dapat dilihat bawah nilai yang sudah melalui proses scaling lebih seimbang dengan nilai dari fitur lainnya.

	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase
0	-0.426715	-0.354665	-0.318393
1	1.682629	-0.091699	-0.034333
2	0.621508	-0.113622	-0.145186
3	-0.447314	-0.365626	-0.311485
4	-0.393756	-0.254379	-0.176363

Gambar 8. Fitur setelah Scaling

D. Analisis menggunakan Algoritma *Machine learning*

Hasil dari pengujian yang telah dilakukan dengan menggunakan beberapa metode *machine learning* seperti berikut , Classification Report Hasil pengujian dengan Algoritma *machine learning* menggunakan train split 80:20.

Tabel 3
Hasil Pengujian

Algoritma	Akurasi	AUC/ROC	Recall	Precision
Random Forest	78,63%	64,88%	37%	64,70%
Gradient Boosting	76,92%	66%	43,33%	56,52%
K Nearest Neighbor	74,35%	59.82%	30,00%	50,00%
Logistic Regression	74,35%	58,73%	26,66%	50,00%
Neural Networks	72,64%	59,77%	33%	45,45%
Decision Tree	71,79%	59,19%	33%	43,47%
Guassian Naive Bayes	55,55%	67,93%	93%	36%

Pada tabel pengujian 3 dengan split 80:20 yang diuji menggunakan seluruh atribut yang sudah melalui tahap preprocessing secara bertahap menghasilkan bahwa akurasi terbaik dimiliki oleh algoritma random forest dengan nilai akurasi sebesar 78,63%, Hasil pengujian yang sudah dilakukan dengan beberapa tahapan preprocessing data.

5. KESIMPULAN

Dari uraian dengan penjelasan dan pembahasan dari materi atau bab sebelumnya serta hasil dari rangkaian pengujian yang dilakukan, maka penulis mengambil kesimpulan bahwa Akurasi menjadi salah satu perhatian penting bagi para peneliti dalam mengembangkan metode diagnosis penyakit. Hasil evaluasi yang telah dilakukan dengan menunjukkan bahwa teknik algoritma Random Dorest yang melalui tahapan preprocessing untuk mengevaluasi dataset penyakit liver. Algoritma Random Forest memiliki performa keakuratan yang diukur dengan akurasi sebesar 78,63%.

6. REFERENSI

Azis, H., Tangguh Admojo, F. dan Susanti, E. (2020) "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, 19(3), hal. 286–294. doi: 10.33633/tc.v19i3.3646.

Falatehan, A. I., Hidayat, N. dan Brata, K. C. (2018) "Sistem Pakar Diagnosis Penyakit Hati Menggunakan Metode Fuzzy Tsukamoto Berbasis Android," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, 2(8), hal. 2373–2381. Tersedia pada: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/1773.

Friedman, J. H. (2001) "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 29(5), hal. 1189–1232. doi: 10.1214/aos/1013203451.

Lin, R. H. (2009) "An intelligent model for liver disease diagnosis," *Artificial Intelligence in Medicine*, 47(1), hal. 53–62. doi: 10.1016/j.artmed.2009.05.005.

Lubis, A. I., Erdiansyah, U. dan Siregar, R. (2022) "Komparasi Akurasi pada Naive Bayes dan Random Forest dalam Klasifikasi Penyakit Liver," *Journal of Computing Engineering, System and Science (CESS)*, 7(1), hal. 81–89.

Prajarini, D. (2016) "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit," 1(3), hal. 1–5.

Rahman, N. T. (2020) "Analisa Algoritma Decision Tree Dan Naïve Bayes Pada Pasien Penyakit Liver," *Jurnal Fasilkom*, 10(2), hal. 144–151. doi: 10.37859/jf.v10i2.2087.

Roihan, A., Sunarya, P. A. dan Rafika, A. S. (2020) "Pemanfaatan *Machine learning* dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian Journal on Computer and Information Technology)*, 5(1), hal. 75–82. doi: 10.31294/ijcit.v5i1.7951.

Utomo, D. P. (2020) "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," 4(April), hal. 437–444. doi: 10.30865/mib.v4i2.2080.

Widodo, H., Rohman, A. dan Sismindari, S. (2019) "Pemanfaatan Tumbuhan Famili Fabaceae untuk Pengobatan Penyakit Liver oleh Pengobat Tradisional Berbagai Etnis di Indonesia," *Media Penelitian dan Pengembangan Kesehatan*, 29(1), hal. 65–88. doi: 10.22435/mpk.v29i1.538.

Wu, Z. *et al.* (2017) "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," *Proceedings - 2017 IEEE International Conference on Computational Science and Engineering and IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, CSE and EUC 2017*, 1, hal. 531–536. doi: 10.1109/CSE-EUC.2017.99.

Zaidan, A. S. A. & R. A. H. & J. k. A. & Z. T. A. & A. A. Z. & B. B. *et al.* (2020) "Role of biological Data Mining and MachineLearningTechniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review," *Journal of Medical Systems*.