PREDIKSI PENYAKIT LIVER MENGGUNAKAN ALGORITMA RANDOM FOREST

Martika Kesuma[1], Sriyanto[2], Sutedi[3]
Institut Informatika dan Bisnis Darmajaya[123]
Jl. ZA Pagar Alam No 93, Bandar Lampung 35142
E-mail : martika.2121219001p@mail.darmajaya.ac.id[1], sriyanto@darmajaya.ac.id[2],
sutedi@darmajaya.ac.id[3]

ABSTRAK

Mendiagnosa sebuah penyakit menggunakan teknologi bukanlah hal yang awam lagi, dengan terus berkembangnya kemajuan dunia kesehatan dapat memanfaatkan teknologi dalam mengambil sebuah keputusan khususnya mendeteksi penyakit liver. Pada dasarnya teknologi yang digunakan dapat sangat membantu para dokter dibandingkan dengan teknik analisis manual konvensional yang selama ini digunakan untuk mendiagnosa penyakit pasien, hati atau yang bisa juga disebut liver merupakan organ tubuh manusia yang fungsinya sangatlah penting, pada penderita penyakit liver fungsi hati akan mulai menurun, menurut WHO (*Word Heart* Organization) tahun 2013 pasien penderita liver di indonesia mencapi 28 juta orang, dari data tersebut membuat penyakit liver disebut sebagai salah satu dari 10 penyakit dengan tingkat kematian yang paling tinggi, akan sangat baik apabila dokter dapat mendeteksi penyakit liver lebih cepat agar pasien dapat segera ditangani oleh dokter. Dari permasalahan diatas yang mendasari penulis melakukan penelitian prediksi penyakit liver. Pada penelitian ini penulis ingin melakukan prediksi penyakit liver menggunakan Algoritma Random Forest. Dalam melakukan pemilihan fitur dan *classifier* yang tepat adalah hal yang paling terpenting untuk peningkatan akurasi dan komputasi dalam prediksi penyakit liver. Peneliti ingin mengetahui apakah algoritma *Random Forest* memiliki nilai akurasi yang tinggi sehingga bisa menjadi landasan untuk menggunakan algoritma *Random Forest* dalam memprediksi penyakit liver. Peneliti menggunakan dataset Liver *Disease Patient* Dataset, dalam tahapan penelitian ini dilakukan beberapa langkah yang dimulai dari melakukan Analisis Data, *Exploratory* Data *Analysis*, *Preprocessing*, Pemodelan Algoritma, dan Visualisasi. Dari tahapan penelitian tersebut telah dapat diketahui hasil prediksi penyakit liver akurasi menggunakan Algortima Random Forest. Dari hasil penelitian yang dilakukan dengan algoritma *Random Forest* didapatkan prediksi dengan nilai akurasi 0.713326 dengan f1 *score* 81%.

Kata kunci : Penyakit Liver, Prediksi Algoritma, *Random Forest, Data Mining*

ABSTRACTS

Diagnosing a disease using technology is no longer commonplace, with the continuous development of advances in the world of health, it can utilize technology in making decisions, especially detecting liver disease. Basically, the technology used can really help doctors compared to conventional manual analysis techniques which have been used to diagnose patient diseases. According to WHO (Word Heart Organization) in 2013, there were 28 million patients with liver disease in Indonesia. From this data, liver disease is referred to as one of the 10 diseases with the highest death rate. It would be very good if doctors could detect liver disease more frequently. quickly so that the patient can be immediately treated by a doctor. From the problems above that underlies the authors to conduct research on the classification of liver disease. In this study the authors wanted to predict liver disease using the Random Forest Algorithm. In selecting the right features and classifiers, the most important thing is to increase accuracy and computation in predicting liver disease. The researcher wants to know whether the Random Forest algorithm has a high accuracy value so that it can be a basis for using the Random Forest algorithm in

predicting liver disease. Researchers used the Liver Disease Patient Dataset, in this research stage several steps were carried out starting from conducting Data Analysis, Exploratory Data Analysis, Preprocessing, Algorithmic Modeling, and Visualization. From this research stage, it can be seen the results of predicting accuracy using the Random Forest Algorithm. From the results of research conducted with the Random Forest algorithm, predictions were obtained with an accuracy value of 0.713326 with an f1 score of 81%.

Keywords : Liver disease, Algorithm prediction, Random Forest, Data Mining.

1. PENDAHULUAN

Dunia kesehatan mengumpulkan sejumlah besar data kesehatan, namun beberapa data kesehatan masih sulit didapatkan. Dalam bidang kesehatan keakuratan dalam prediksi sebuah penyakit memerlukan keputusan yang efektif dalam mengambil keputusan dan keakuratan dalam prediksi suatu penyakit sangat penting. Hati merupakan salah satu organ yang cukup besar dan penting pada tubuh kita. Bagian yang penting pada hati ini terdiri dari hepatosit, yang merupakan sel epitel dengan konfigurasi yang unik. Pada dasarnya hati ini merupakan kelenjar eksorin, oleh karena mensekresi cairan empedu yang dialirkan kedalam duodenum. Selain itu juga merupakan kelenjar endokrin dan penyaring darah[1].

Dalam organ ini pun terdapat penyakit hati atau biasa disebut liver, liver memiliki beragam fungsi penting, antara lain membersihkan darah dari senyawa berbahaya. Selain itu, hati juga memproduksi protein yang berperan penting dalam proses pembekuan darah[2]. Hati dapat memperbaiki sel-selnya yang rusak. Namun, pada penderita penyakit liver, sel-sel hati yang rusak cukup banyak sehingga fungsinya terganggu. Biasanya, fungsi hati akan mulai menurun ketika sel-sel hati yang rusak mencapai 75%. Penurunan fungsi hati umumnya terjadi secara bertahap. Kerusakan yang timbul akibat penurunan fungsi hati akan mengikuti perkembangan penyakit yang mendasarinya. 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[3]. 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 tersebut membuat penyakit liver disebut sebagai salah satu dari 10 penyakit dengan tingkat kematian yang paling tinggi sehingga angka kematian setiap tahun semakin meningkat[4].

Dalam pemaparan latar belakang diatas perlu adanya pengembangan penelitian dalam masalah sistem pendukung keputusan medis yang cerdas dan efektif untuk membantu para dokter. Pemilihan fitur dan *classifier* yang tepat adalah hal terpenting dalam peningkatan akurasi dan komputasi dalam prediksi penyakit liver salah satunya yaitu *Random Forest* yang digunakan untuk mengklasifikasikan data dalam jumlah besar. Selain itu harus diketahui sejauh mana tingkat akurasi yang didapat dari model. Dengan cara memprediksi algoritma klasifikasi yang ditentukan, dengan *Random Forest* untuk pengambilan keputusan dalam menentukan prediksi peyakit liver pada pasien yang terkena atau tidaknya terhadap penyakit liver tersebut[5].

2. METODE PENELITIAN

Dalam penelitian ini terdiri dalam beberapa langkah pengerjaan yaitu tahapan *analyzing* dataset kemudian dilanjutkan dengan deskripsi dataset menggunakan EDA (*Exploratory* Data *Analysis*) setelah itu tahapan *preprocessing* data salah satunya yaitu pengecekan *missing* pada dataset kemudian dilanjutkan pada tahapan pemodelan *machine learning* untuk menentukan hasil *score* pada setiap algoritma yang dibandingkan. Dalam tahapan ini akan dilakukan *splitting* dataset untuk data *training*

P-ISSN: 2337-8344
E-ISSN: 2623-1247

dan data *testing* dan menggunakan 10-*fold cross validation* dalam pemodelannya yang kemudian membuat prediksi menggunakan algoritma *Random Forest* setelah itu tahapan visualisasi dan hasil.

3. HASIL DAN PEMBAHASAN

3.1 *Analyzing* Dataset

Dalam tahapan ini dilakukan deskripsi pada dataset untuk melihat nilai pada setiap atribut yang digunakan. Berikut adalah hasilnya :

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Al
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	
mean	44.746141	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190	3.141852	
std	16.189833	6.209522	2.808498	242.937989	182.620356	288.918529	1.085451	0.795519	
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	
25%	33.000000	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000	2.600000	
50%	45.000000	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000	3.100000	
75%	58.000000	2.600000	1.300000	298.000000	60.500000	87.000000	7.200000	3.800000	
max	90.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	

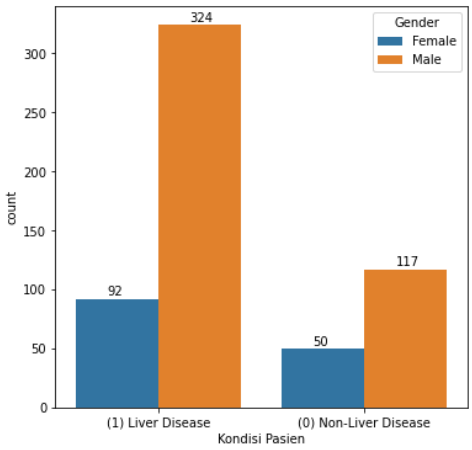
Setelah mendeskripsikan data dilanjutkan dengan mencetak bentuk data seperti hasil berikut :

```
(583, 11)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 583 entries, 0 to 582
Data columns (total 11 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Age                         583 non-null    int64
 1   Gender                      583 non-null    object
 2   Total_Bilirubin             583 non-null    float64
 3   Direct_Bilirubin            583 non-null    float64
 4   Alkaline_Phosphotase        583 non-null    int64
 5   Alamine_Aminotransferase    583 non-null    int64
 6   Aspartate_Aminotransferase  583 non-null    int64
 7   Total_Protiens              583 non-null    float64
 8   Albumin                     583 non-null    float64
 9   Albumin_and_Globulin_Ratio  579 non-null    float64
 10  Result                      583 non-null    int64
dtypes: float64(5), int64(5), object(1)
memory usage: 50.2+ KB
```

Dapat dilihat setiap atribut yang digunakan ditampilkan tipe data yang digunakan, terdapat 11 atribut pada dataset yang digunakan dengan 5 tipe data yang berbeda.

3.2 *Exploratory* Data *Analysis*

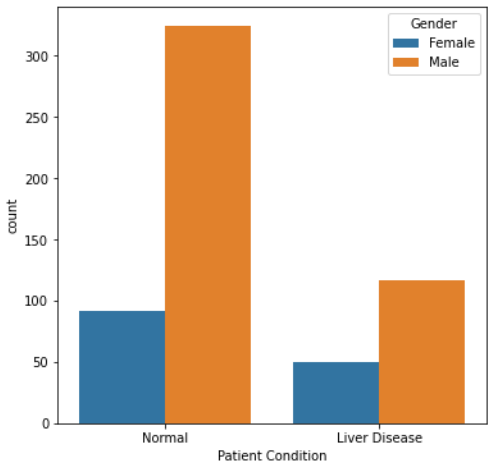
Exploratory Data *Analysis* adalah suatu proses uji investigasi awal yang bertujuan untuk mengidentifikasi pola, menemukan anomali, menguji hipotesis dan memeriksa asumsi. Dengan melakukan EDA, akan sangat berguna dalam mendeteksi kesalahan dari awal, dapat mengidentifikasi *outlier*, mengetahui hubungan antar data serta dapat menggali faktor-faktor penting dari data. Dalam proses EDA akan dilakukan analisis data dengan menampilkan model visualisasi dalam dataset yang digunakan. Dalam tahapan pertama akan memvisualisasikan jumlah pasien yang di diagnosis mengalami penyakit hati. Berikut adalah hasil visualisasinya.



Gambar 1. Visualisasi jumlah pasien terdiagnosis penyakit hati

Yang selanjutnya yaitu memvisualisasikan data dengan penyakit hati beserta jenis kelamin seperti berikut :

```
Text(0.5, 0, 'Patient Condition')
```

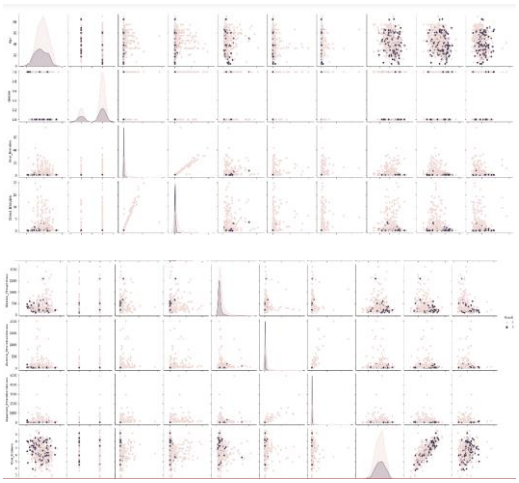


Gambar 2. Visualisasi data dengan penyakit hati dan jenis kelamin

Dapat dilihat dari hasil visualisasi diatas jumlah bahwa jumlah *gender* dengan kondisi terdiagnosis penyakit hati lebih banyak *male* dari pada *female* dalam dataset yang tersedia. Pada tiap-tiap atribut dilakukan pemodelan untuk pengkodean atribut untuk mem*ploting corelasi* antara tiap-tiap atribut yang terdapat pada dataset. Berikut adalah hasil analisis pola

P-ISSN: 2337-8344
E-ISSN: 2623-1247

yang dilakukan dengan *Exploratory* Data *Analysis* seperti gambar berikut :



Gambar 3. Data *Analysis*

3.3 *Preprocessing*

Data *preprocessing* adalah proses mengubah data mentah ke dalam bentuk yang lebih mudah dipahami. Proses ini dilakukan untuk memperbaiki kesalahan pada data mentah yang tidak lengkap dan memiliki format yang tidak teratur. Dalam tahapan *preprocessing* dalam penelitian ini adalah dilakukan pengecekan pada data yang hilang atau biasa disebut *missing values* yang terdapat dalam dataset. Tahap pertama pengecekan *missing* data pada masing-masing atribut seperti hasil berikut :

```
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
Result                      0
dtype: int64
```

Gambar 4. *Missing Value*

Dapat dilihat pada gambar diatas terdapat *missing* dalam atribut *albumin and globulin ratio* dengan 4 *missing value*. Setelah pengecekan *missing value* dilanjut dengan pengecekan pada baris pada data yang *missing* seperti hasil berikut :

Gambar 5. *Missing Values* Pada Baris

Setelah terlihat data yang *missing* pada baris dataset maka kita lihat korelasi rasio antara *albumin and globulin ratio* dengan kolom lainnya kemudian dilakukan perhitungan rata-rata untuk menimpa data yang kosong sehingga data dapat bersih seperti hasil berikut :

```
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
Result                      0
dtype: int64
```

Gambar 6. Hasil *preprocessing* pada data *missing*

Dapat dilihat data pada *albumin and globulin ratio* menjadi 0 yang artinya tidak ada data yang kosong seperti sebelumnya. Setelah tahapan *missing value* pada dataset dilakukan *featr selection* pada tiga algoritma yang akan digunakan untuk perbandingan sebagai pemodelan data.

3.4 Pemodelan Algoritma

Pada tahapan pemodelan setelah tidak ada data yang *missing* agar tidak terjadi kekosongan pada data yang akan diolah, selanjutnya adalah tahapan pada *splitting* dataset sebagai data yang akan di latih dan juga data untuk *testing*. Pada tahapan *splitting* data menggunakan 10-*fold cross validation* dan membuat komparasi perbandingan pada tiga algoritma pembelajaran *learning*. Setelah pemodelan tiga algoritma maka akan dilakukan inisiasi pada model untuk melihat performa dari masing-masing model

P-ISSN: 2337-8344
E-ISSN: 2623-1247

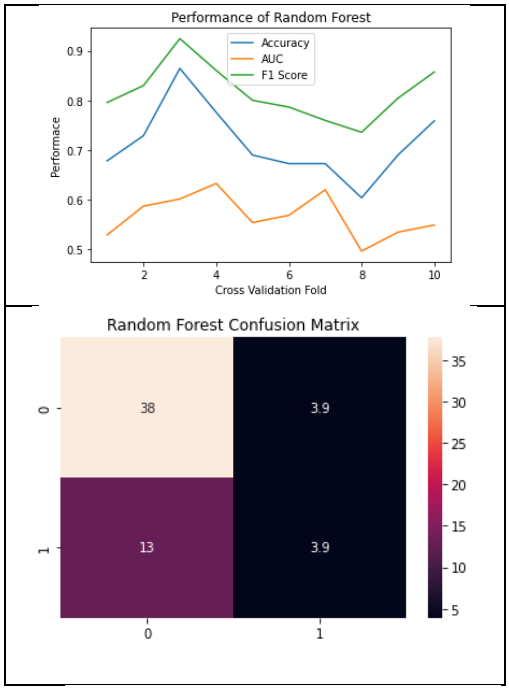
dan juga untuk tahapan klasifikasi hingga terbentuk 10 *fold* berikut :

```
1 Number of fold
2 Number of fold
3 Number of fold
4 Number of fold
5 Number of fold
6 Number of fold
7 Number of fold
8 Number of fold
9 Number of fold
10 Number of fold
```

Gambar 7. 10-*fold cross validation*

3.5 Hasil Analisis Performa

Pada hasil analisis performa yaitu tahap akhir dari hasil prediksi pada algoritma yang telah dibuat modelnya. Pada tahap ini akan ditampilkan model visualisasi dan *result* pada algoritmanya. Pada tahapan ini akan ditampilkan f1 *score*, performa dari algoritma, *cross validation fold* dan performa. Berikut adalah hasilnya.



Gambar 8. Model Dan *Confusion Matrix Random Forest*

Pada hasil pemodelan algoritma diolah dapat dilihat bahwa algoritma *Random Forest* memiliki hasil yang baik. Berikut adalah hasil visualiasi dengan menggunakan data*frame* berdasarkan hasil visualiasi pada gambar diatas.

	Algorithms	Accuracy	Confusion Matrix	AUC	F1 Score
0	Random Forest	0.69962	[[37.5, 4.1], [13.4, 3.3]]	0.550644	0.808319

Gambar 11. Hasil Prediksi Algoritma *Random Forest*

Pada Hasil Prediksi menggunakan algoritma dapat dilihat bahwa algoritma *Random forerst* mendapatkan hasil yang cukup baik untuk melakukan sebuah prediksi pada penyakit liver pada pemodelan pembelajaran *learning* dengan f1 *score* 81%.

4. KESIMPULAN DAN SARAN

Penelitian ini bertujuan untuk mengetahui hasil Prediksi pada Algoritma *Random Forest* yang memiliki nilai akurasi cukup baik untuk melakukan sebuah prediksi pada penyakit liver. Pemodelan Algoritma yang dihasilkan dengan algoritma *Random Forest* memiliki nilai *accuracy* 0.713326 dan f1 *score* 81%.

Pada penelitian selanjutnya dapat ditambahkan atau dikombinasikan dengan algoritma lain untuk meningkatkan akurasi. Karena jika melihat kinerja performa yang dibuat dari algoritma masih dapat ditingkatkan untuk penelitian selanjutnya.

DAFTAR PUSTAKA

[1] W. Erawati, "Vol. XII No. 2, September 2015 Jurnal Techno Nusa Mandiri," *Techno Nusa Mandiri*, vol. XII, no. 2, pp. 21–26, 2015.

[2] Elly Pusporani, Siti Qomariyah, and Irhamah, "Klasifikasi Pasien Penderita Penyakit Liver," vol. 2, no. March, 2019.

[3] R. Amin, R. Yasmin, S. Ruhi, M. H. Rahman, and M. S. Reza, "Prediction of chronic liver disease patients using integrated projection

based statistical feature extraction with machine learning algorithms," *Informatics Med. Unlocked*, vol. 36, no. December 2022, p. 101155, 2023, doi: 10.1016/j.imu.2022.101155.

[4] E. F. 4) F. Lia Dwi Cahyanti1), Fajar Sarasati2), Widi Astuti 3), "KLASIFIKASI DATA MINING DENGAN ALGORITMA MACHINE LARNING UNTUK PREDIKSI PENYAKIT LIVER," vol. 14, no. 2, 2023.

[5] M. R. F. Rizki, "Perbandingan Algoritma Klasifikasi Untuk Prediksi Penyakit Liver," *Reputasi J. Rekayasa Perangkat Lunak*, vol. 1, no. 2, pp. 82–88, 2020, doi: 10.31294/reputasi.v1i2.109.