JTIM : Jurnal Teknologi Informasi dan Multimedia

p-ISSN : 2715-2529
e-ISSN : 2684-9151
https://journal.sekawan-org.id/index.php/jtim/

Analisis Pengaruh Komposisi *Data Training* dan *Data Testing* pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver

Baiq Nurul Azmi[1*], Arief Hermawan[2], dan Donny Avianto[3]

[1] Magister Teknologi Informasi, Universitas Teknologi Yogyakarta ; 6210211006.baiq@student.uty.ac.id
[2] Magister Teknologi Informasi, Universitas Teknologi Yogyakarta; ariefdb@uty.ac.id
[3] Informatika, Universitas Teknologi Yogyakarta; donny@uty.ac.id
* Korespondensi: 6210211006.baiq@student.uty.ac.id

Sitasi: Azmi, B. N.; Hermawan, A; Avianto, D. (**2023**). Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver . JTIM: Jurnal Teknologi Informasi Dan Multimedia, 4(4), 281-290
https://doi.org/10.35746/jtim.v4i4.298

Abstract: Liver disease is one of the diseases that is difficult to detect and becomes the largest contributor to deaths because it is considered a silent killer without symptoms. Liver disease can be detected based on abnormalities in the number of contents in the human body. The Indian Liver Patient Dataset (ILPD) dataset has many variables related to content in the body of liver patient data which are used as parameters in the classification of liver disease patients. Previous studies have shown that only two variables influence the ILPD dataset. The purpose of this study is to examine the use of the Principal Component Analysis (PCA) method to determine the optimal number of features in the context of classification of liver disease and examine the percentage distribution of data training and data testing which produces the best accuracy. The ILPD dataset was obtained from the UCI Machine Learning website with a total of 583 rows of data and 11 features. The percentage of training data and testing data used is 50%:50%, 60%:40%, 70%:30%, 73%:27%, 75%:25%, 80%:20%, 83%:17%, 85%:15% and 90%:10%. The use of different training and testing data percentages serves to find the best accuracy. The PCA result feature is used as input for the Decision Tree C4.5 classification algorithm. The experimental results show that using the training and testing data distribution percentage of 90%:10% and after the application of PCA produces the highest accuracy, namely 78.40% which is obtained for the number of PCA components n = 8.

Keywords: PCA, Decision Tree, Liver Disease, Training, Testing

Abstrak: Penyakit liver merupakan salah satu penyakit yang sulit dideteksi dan menjadi penyumbang kematian terbesar karena dianggap sebagai pembunuh diam-diam tanpa gejala. Penyakit liver dapat dideteksi berdasarkan kelainan jumlah kandungan-kandungan yang ada di dalam tubuh manusia. Dataset Indian Liver Patient Dataset (ILPD) memiliki banyak variabel terkait kandungan dalam tubuh data pasien liver yang dijadikan parameter dalam klasifikasi pasien penyakit liver. Penelitian sebelumnya menunjukkan bahwa variabel yang berpengaruh pada dataset ILPD hanya dua yaitu variabel Age dan SGPT_AA. Tujuan penelitian ini mengkaji penggunaan metode Principal Component Analysis (PCA) untuk menentukan jumlah fitur yang paling optimal dalam konteks klasifikasi penyakit liver dan mengkaji pembagian persentase data training dan data testing yang dapat menghasilkan akurasi terbaik. Data Indian Liver Patient Dataset (ILPD) diperoleh dari situs UCI Machine Learning Repository dengan total data yaitu 583 baris data dan 11 fitur. Persentase pembagian data training dan data testing yang digunakan adalah 50%:50%, 60%:40%, 70%:30%, 73%:27%, 75%:25%, 80%:20%, 83%:17%, 85%:15% dan 90%:10%.

Penggunaan beberapa persentase *data training* dan *data testing* yang berbeda berfungsi untuk menemukan akurasi terbaik. Fitur hasil Principal Component Analysis (PCA) digunakan sebagai input untuk algoritma klasifikasi Decision Tree C4.5. Hasil eksperimen menunjukkan dengan penggunaan persentase pembagian *data training* dan *data testing* 90%:10% serta setelah penerapan PCA menghasilkan akurasi tertinggi yaitu 78.40% yang didapatkan pada jumlah komponen PCA n=8.

Kata kunci: PCA, Decision Tree, Penyakit Liver, Training, Testing

1. Pendahuluan

Hati atau liver adalah organ terpenting dan terbesar seorang manusia yang digunakan untuk membentuk dan mengeluarkan empedu sebagai detoksifikasi racun. Hati dapat terkena penyakit yang disebabkan oleh peradangan pada organ hati. Penyebab terjadinya peradangan pada hati yaitu kelainan organ hati sejak lahir dan juga pola hidup yang tidak sehat seperti ketergantungan alkohol, senyawa kimia, dan obat-obatan [1].

Penyakit liver merupakan salah satu penyakit yang sulit dideteksi dan menjadi penyumbang kematian karena dianggap sebagai pembunuh diam-diam tanpa gejala [2]. Hasil analisis data kematian yang dirilis oleh British Liver Trust pada tahun 2019 mengungkapkan bahwa penyakit hati adalah penyebab kematian terbesar di Inggris dan Wales pada rentang usia 35-49 tahun, dimana pada penelitian tersebut juga diperkirakan bahwa penyakit hati kan menggeser penyakit jantung sebagai penyebab terbesar kematian dini dalam beberpa tahun mendatang [3].

Pemeriksaan kesehatan secara berkala diperlukan untuk mendeteksi penyakit liver sejak dini, sehingga dapat di tangani lebih cepat dan tingkat kematian karena penyakit liver juga dapat menurun. Penyakit liver dapat dideteksi berdasarkan kelainan jumlah kandungan-kandungan dalam tubuh manusia. Dataset Indian Liver Patient Dataset (ILPD) memiliki banyak variabel terkait kandungan dalam tubuh data pasien liver yang dijadikan parameter dalam klasifikasi pasien penyakit liver.

Proses klasifikasi penyakit liver dapat dilakukan dengan berbagai algoritma klasifikasi. Penelitian terkait klasifikasi penyakit liver dengan dataset ILPD sebelumnya pernah dilakukan oleh [4] yang bertujuan untuk mengetahui variabel yang paling berpengaruh dalam klasifikasi penyakit liver menggunakan algoritma C4.5. Hasil penelitian menunjukkan akurasi yang diperoleh sebesar 72.67% dan hanya dua variabel yaitu Age dan Sgpt_AA yang menjadi variabel paling berpengaruh dalam klasifikasi penyakit liver.

Penelitian lainnya yang membahas klasifikasi penyakit liver menggunakan dataset ILPD yaitu penelitian pada tahun 2018 yang melakukan pendekatan seleksi fitur dengan proses klasifikasi menggunakan algoritma Cart dan Ripper dengan tools WEKA. Seleksi Atribut menyebabkan hanya 7 atribut yang terpilih, dan menghasilkan akurasi sebesar 70% dengan recall 84% dan precision sebesar 58% [5]. Penelitian dilakukan juga oleh [2] yang mengkomparasi algoritma Decision Tree dan Naïve Bayes pada pasien penyakit liver, setelah dilakukan pengujian dengan split validation dan cross validation, didapatkan akurasi untuk algoritma Decision Tree sebesar 70.29% dan algoritma Naïve Bayes sebesar 67.05%.

Penelitian sebelumnya membahas tentang *split* data dapat mempengaruhi nilai akurasi dari suatu pengolahan data, penelitian tersebut melakukan pengolahan data produk dengan mengkombinasikan metode K-Means dengan Decision Tree, serta menggunakan pembagian *data training* dan *data testing* sebanyak 70%:30%, 80%:20%, dan 90%:10%, yang menghasilkan nilai akurasi tertinggi yaitu 98.77%, didapatkan pada persentase 90%:10% [6]. Penelitian lainnya yang dilakukan oleh [7] juga menggunakan beberapa pembagian *data training* dan *data testing* yang berbeda pada klasifikasi masa

studi mahasiswa dengan algoritma naïve bayes dan bayesian network, yaitu dengan pembagian *data training* dan *data testing* berturut-turut sebesar 30%:70%, 40%:60%, 50%:50%, 60%:40%, 70%:30%, 80%:20%, dan 90%:10% yang menghasilkan kinerja terbaik didapatkan pada kedua algoritma menggunakan persentase *data traning* dan *data testing* 90%:10% dengan akurasi 80%. Hasil dari penelitian-penelitian sebelumnya menunjukkan bahwa penentuan jumlah persentase *data training* dan *data testing* yang tepat berpengaruh pada kinerja hasil klasifikasi.

Faktor persentase pembagian jumlah *data training* dan *data testing* untuk klasifikasi penderita penyakit liver dapat mempengaruhi kinerja model klasifikasi. Penggunaan seluruh fitur dalam proses klasifikasi juga dapat mempengaruhi kinerja model klasifikasi, baik meningkatkan atau menurunkan performa klasifikasi dalam hal akurasi dan juga waktu komputasi yang diperlukan, sehingga pendekatan yang dapat dilakukan yaitu dengan memilih persentase *data training* dan data uji yang tepat serta melakukan reduksi dimensi.

Penelitian ini mengolah dataset ILPD berdasarkan parameter algoritma klasifikasi Decision Tree dari penelitian sebelumnya yang dilakukan oleh [4] dengan upaya pendekatan percobaan menggunakan persentase *data training* dan *data testing* terbaik serta menerapkan metode reduksi dimensi Principal Component Analysis (PCA). Penelitian ini bertujuan untuk memperbaiki hasil akurasi dari penelitian-penelitian sebelumnya dengan algoritma decision tree, selain itu untuk mengetahui persentase *data training* dan *data testing* terbaik untuk dataset ILPD serta untuk menganalisis pengaruh metode reduksi dimensi PCA.

2. Dataset dan Metode

2.1 Dataset

Penelitian ini menggunakan data sekunder dari situs koleksi dataset UCI Machine Learning Repository yang berjudul Indian Liver Patient Dataset (ILPD) [8]. Dataset ILPD berisi tentang data penderita penyakit hati dengan jumlah data yaitu 11 atribut dan 583 baris data. Tabel 1 disajikan dataset ILPD dengan keterangan 11 atribut yaitu *Age, Gender, TB (Total Bilirubin), DB (Direct Bilirubin), Alkphos Alkaline Phosphotase, Sgpt Alamine Aminotransferase, Sgot Aspartate Aminotransferase, TP Total Protiens, ALB Albumin, A/G Ratio Albumin and Globulin Ratio, Class (Label Positive / Negative Liver).*

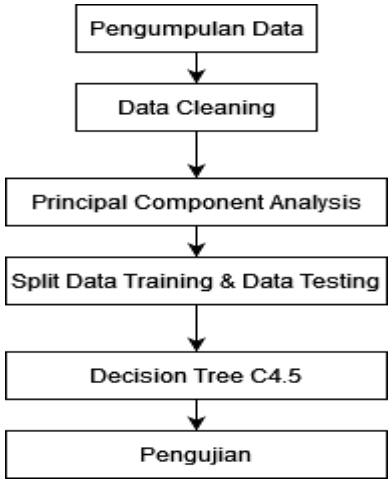
Tabel 1. Data Indian Liver Patient Dataset (ILPD)

No	Age	Gender	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	A/G	Class
1	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
2	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
3	62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
4	58	Male	1	0.4	182	14	20	6.8	3.4	1	1
5	72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
6	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
7	26	Female	0.9	0.2	154	16	12	7	3.5	1	1
8	29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
…	…	…	…	…	…	…	…	…	…	…	…
583	38	Male	1	0.3	216	21	24	7.3	4.4	1.5	2

2.2 Metode

Penelitian ini dilaksanakan untuk meningkatkan hasil kinerja pada penelitian terdahulu oleh [4] dalam klasifikasi penyakit liver dengan akurasi 72.67%, dimana pada penelitian ini dilakukan klasfikasi penyakit liver menggunakan reduksi dimensi Principal

Component Analysis dan algoritma Decision Tree C4.5 dengan ratio split data yang berbeda pada setiap percobaan. Tahapan penelitian dapat dilihat pada Gambar 1.



Gambar. 1. Tahapan Penelitian

2.2.1 Pengumpulan Data

Data yang digunakan adalah data penyakit liver yang diperoleh dari situs UCI Machine Learning. Data berjudul Indiam Liver Patient Dataset (ILPD). Data ini dapat diakses dan digunakan dengan bebas. Dataset dapat dilihat pada Tabel 1.

2.2.2 Data Cleaning

Data yang didapatkan awalnya akan dibersihkan terlebih dahulu pada proses *data cleaning*. *Data cleaning* adalah proses mendeteksi, memperbaiki atau bahkan menghapus catatan, tabel, dan database yang salah atau tidak akurat [9]. Pada tahap ini dilakukan pengecekan terhadap keberadaan data kosong (*missing value*), setelah itu dilakukan pengisian terhadap data yang kosong.

2.2.3 Principal Component Analysis

Principal Component Analysis (PCA) merupakan metode untuk mereduksi dimensi atribut dalam dataset, sehingga nilai yang terbentuk sangat berbeda dengan bentuk aslinya [10]. Metode PCA digunakan untuk meringkas struktur dari dataset dengan dimensi yang banyak sehingga memiliki jumlah variabel yang lebih kecil [11]. Kegunaan PCA selain untuk mereduksi dimensi yaitu dapat digunakan sebagai metode untuk menguji apakah setiap variabel dalam dataset saling terkait atau tidak terkait sama sekali.

2.2.4 Pembagian *Data Training* dan *Data Testing*

Data training adalah data yang benar-benar ada sebelumnya sesuai dengan fakta sedangkan *data testing* adalah data yang digunakan untuk mengukur sejauh mana pengklasifikasi berhasil mengklasifikasikan dengan benar [12]. Pembagian jumlah *data training* dan *data testing* adalah salah satu faktor yang menentukan akurasi, sehingga kesalahan dalam menentukan komposisi kedua tipe data tersebut akan mempengaruhi nilai akurasi dan presisi yang diperoleh [13]. *Data training* dan *data testing* biasa digunakan pada machine learning. Mesin diberikan sekelompok dataset untuk dipelajari dan disebut *data training*, kemudian hasil pembelajaran selanjutnya akan digunakan untuk mengolah dataset baru yang disebut *data testing* [14]. Penelitian ini melakukan percobaan klasifikasi dengan sembilan persentase *data training* dan *data testing* yang berbeda, yaitu dengan persentase 50%:50%, 60%:40%, 70%:30%, 73%:27%, 75%:25%, 80%:20% , 83%:17%,

85%:15%, dan 90%:10%. Pembagian jumlah *data training* dan *data testing* pada sembilan persentase yang berbeda disajikan pada Tabel 2.

Tabel 2. Jumlah *data training* dan *data testing*

Persentase (*Training:Testing*)	Jumlah *Data training*	Jumlah *Data testing*
50%:50%	292	291
60%:40%	394	234
70%:30%	408	175
73%:27%	425	158
75%:25%	437	146
80%:20%	466	117
83%:17%	483	100
85%:15%	495	88
90%:10%	523	60

2.2.5 Decision Tree C4.5

Decision Tree merupakan teknik model prediksi yang dapat digunakan untuk klasifikasi dan prediksi tugas [15]. Decision Tree menggunakan teknik "*divide and conquer*" untuk membagi ruang pencarian masalah menjadi himpunan masalah [16].

Ada dua jenis algoritma Decision Tree yang terkenal, yaitu C4.5 dan Random Forest. Pada penelitian ini akan menggunakan algoritma C4.5. Algoritma C4.5 adalah algoritma yang dikembangkan dari algoritma ID3 dengan beragam peningkatan. Algoritma C4.5 merupakan struktur pohon keputusan dimana terdapat simpul yang mendiskripsikan atribut – atribut, setiap cabang menggambarkan hasil dari atribut yang duji, dan setiap daun menggambarkan kelas. Beberapa peningkatan ini diantaranya adalah, penanganan atribut-atribut numerik, *missing value* dan *noise* pada dataset, dan aturan-aturan yang dihasilkan dari model pohon yang terbentuk [17].

Langkah kerja algoritma C4.5 yaitu menghitung nilai entropy, nilai *information gain*, nilai *split info*, dan nilai *gain ratio* dijelaskan sebagai berikut [18]:

1. Menghitung entropi setiap atribut, menggunakan Persamaan 1.

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i log_2 p_i \quad (1)$$

dimana

S = Himpunan Kasus

n = Jumlah partisi S

p_i = Probabilitas yang didapat dari jumlah kelas dibagi total kasus

2. Menghitung *information gain* dari setiap atribut, menggunakan Persamaan 2.

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si) \quad (2)$$

dimana

S = Himpunan Kasus

A = Atribut

n = Jumlah partisi S

|Si| = Jumlah partisi ke-i

|S| = Jumlah kasus dalam S

3. Menghitung *split gain* dari setiap atribut, menggunakan Persamaan 3.

$$Split\ Infor\ (S,A) = \sum_{i=1}^{n} \frac{|Si|}{|S|}\ log_2 \frac{|Si|}{|S|} \quad (3)$$

dimana

S = Ruang sampel data yang digunakan untuk training

A = Atribut

|Si| = Jumlah sampel atribut i

4. Menghitung *gain ratio* dari setiap atribut menggunakan Persamaan 4.

$$Gain\ ratio\ (S,A) = \frac{Information\ Gain\ (S,A)}{Split\ Info\ (S,A)} \quad (4)$$

2.2.6 Pengujian Hasil

Pengujian dilakukan untuk mengukur model klasfikasi yang terbentuk . Pengujian yang dilakukan pada penelitian ini yaitu dengan mencari nilai akurasi. Akurasi akan memberikan gambaran seberapa dekat nilai aktual dengan nilai hasil prediksi (Persamaan 5).

$$Akurasi = \frac{tp + tn}{tp + fn + fp\ + tn} \quad (5)$$

dimana :

a. TP *(True Positif)* merupakan banyaknya data di kelas aktualnya positif dan kelas prediksi positif.
b. FN *(False Negative)* merupakan banyaknya data di kelas aktualnya positif dan kelas prediksi negative.
c. FP *(False Positive)* merupakan banyaknya data di kelas aktualnya negatif dan kelas prediksi positif.
d. TN *(True Negative)* merupakan banyaknya data di kelas aktualnya negatif dan kelas prediksi negatif.

3. Hasil

3.1. Preprocessing Data

Dataset ILPD terdiri dari 583 baris data dan 11 atribut, seperti yang ditunjukkan pada Tabel 1. Tahap *preprocessing* dilakukan pertama kali setelah data dikumpulkan. Tahapan *preprocessing* yang dilakukan pada penelitian ini yaitu *data cleaning*, dan reduksi dimensi.

Tahap *data cleaning* yang dilakukan yaitu menangani *missing value* pada dataset ILPD. *Data cleaning* yang dilakukan pada dataset ILPD berupa pengisian data yang kosong pada suatu atribut dengan median dari keseluruhan nilai pada atribut tersebut. Pada dataset ILPD ditemukan 4 data kosong tepatnya di atribut A/G, lalu data tersebut diisi dengan median dari nilai atribut A/G, sehingga *missing value* sudah di tangani dan jumlah data tetap 583 baris.

Tahap selanjutnya yaitu melakukan reduksi dimensi dengan menggunakan metode PCA. Reduksi dimensi dengan metode PCA dilakukan untuk megetahui atribut-atribut yang terdapat pada dataset saling berhubungan atau tidak. Atribut pada dataset ILPD berjumlah 11, dimana 1 atribut sebagai label dan 10 atribut lainnya yang menjadi fitur

untuk klasifikasi dan di reduksi terlebih dahulu dengan metode PCA. Penelitian ini mencoba mereduksi atribut dataset ILPD mulai dari 1 komponen hingga 10 komponen.

3.2. Pemrosesan Data

Pemrosesan data dilakukan setelah tahap *preprocessing* data. Sebelum proses klasifikasi, dataset harus dibagi terlebih dahulu menjadi *data training* dan *data testing*. *Data training* ajan digunakan untuk membentuk model Decision Tree, sedangkan *data testing* dgunakan untuk menguji performa dari algoritma Decision Tree.

Pembagian persentase *data training* dan *data testing* yang digunakan pada penelitian ini adalah 50%:50%, 60%:40%, 70%:30%, 73%:27%, 75%:25%, 80%:20%, 83%:17%, 85%:15%, dan 90%:10%. Persentase-persentase tersebut dicoba untuk mencari jumlah rasio terbaik untuk *data training* dan *data testing*, untuk mendapatkan hasil akurasi terbaik.

Setelah *data training* dan *data testing* di bagi, maka proses selanjutnya adalah melakukan proses klasifikasi menggunakan algoritma Decision Tree. Algoritma Decision Tree dibangun dengan random_state=1, max_depth =. Diharapkan dengan eksperimen menggunakan parameter tersebut dapat menghasilkan akurasi terbaik.

3.3. Pengujian

Penelitian ini menggunakan skema pengujian dengan persentase *data training* dan *data testing* sebesar 50%:50%, 60%:40%, 70%:30%, 73%:27%, 75%:25%, 80%:20%, 83%:17%, 85%:15%, dan 90%:10% dengan algoritma Decision Tree. Penelitian ini juga melakukan percobaan pengujian tanpa metode reduksi PCA dan menerapkan metode reduksi dimensi PCA dengan jumlah komponen dari 1 hingga 10 komponen, yang akan dilakukan berulang-ulang untuk masing-masing persentase pembagian *data training* dan *data testing*.

Hasil pengujian untuk persentase *data training* dan *data testing* sebesar 50%:50%, 60%:40%, 70%:30%, 73%:27%, 75%:25%, 80%:20% , 83%:17%, 85%:15%, dan 90%:10% pada algoritma Decision Tree tanpa PCA disajikan pada Tabel 3.

Tabel 3. Hasil Pengujian Decision Tree tanpa PCA

Split Data	Akurasi
50%:50%	70.5%
60%:40%	70.1%
70%:30%	70.3%
73%:27%	60.8%
75%:25%	72.6%
80%:20%	66.6%
83%:17%	70.0%
85%:15%	76.0%
90%:10%	76.1%

Tabel 3 menunjukkan hasil pengujian algoritma Decision Tree menggunakan sembilan persentase data *training* dan data *testing* yang berbeda tanpa penerapan metode reduksi dimensi PCA, didapatkan hasil akurasi terendah pada persentase 73%:27% yaitu 60.8%, sedangkan hasil akurasi tertinggi didapatkan pada persentase 90%:10% yaitu sebesar 76.1%.

Hasil pengujian untuk persentase *data training* dan *data testing* sebesar 50%:50%, 60%:40%, 70%:30%, 73%:27%, 75%:25%, 80%:20%, 83%:17%, 85%:15%, dan 90%:10% pada algoritma Decision Tree dengan metode reduksi dimensi PCA disajikan pada Tabel 4.

Tabel 4. Hasil Pengujian Decision Tree dengan PCA

n PCA	50:50	60:40	70:30	73:27	75:25	80:20	83:17	85:15	90:10
1	**70.89%**	**72.22%**	**71.42%**	**69.62%**	**70.54%**	70.94%	73.00%	76.13%	76.27%
2	70.89%	72.22%	71.42%	69.62%	70.54%	71.79%	73.00%	76.13%	76.27%
3	70.89%	72.22%	71.42%	69.62%	70.54%	71.79%	73.00%	76.13%	76.27%
4	70.89%	71.36%	71.42%	69.62%	70.54%	71.79%	68.00%	76.13%	76.27%
5	70.89%	71.36%	70.85%	69.62%	70.54%	71.79%	74.00%	75.00%	76.27%
6	70.89%	71.36%	70.85%	69.62%	70.54%	71.79%	74.00%	75.00%	76.27%
7	70.89%	71.36%	70.85%	69.62%	70.54%	**72.64%**	74.00%	75.00%	76.27%
8	70.89%	71.36%	70.85%	69.62%	70.54%	72.64%	**75.00%**	**77.96%**	**78.40%**
9	70.89%	71.36%	70.85%	69.62%	70.54%	72.64%	75.00%	77.96%	78.40%
10	70.89%	71.36%	70.85%	69.62%	70.54%	72.64%	75.00%	77.96%	78.40%

Tabel 4 menunjukkan hasil dari pengujian algoritma Decision Tree dengan PCA menggunakan *data training* dan *data testing* sebesar 50%:50% didapatkan hasil akurasi sama pada setiap jumlah komponen PCA dari 1 hingga 10 yaitu 70.89%. Pengujian dengan perbandingan 60%:40% mendapatkan hasil akurasi tertinggi yaitu 72.22% pada komponen PCA n=1 sampai n=3. Sedangkan untuk pengujian dengan perbandingan 70%:30%, hasil akurasi tertinggi yaitu 71.42% didapatkan pada jumlah komponen PCA n=1 sampai n=4. Pengujian dengan perbandingan 73%:27% didapatkan hasil akurasi sama pada setiap jumlah komponen PCA dari 1 hingga 10 yaitu 69.62%. Pengujian dengan perbandingan sebesar 75%:25%, didapatkan hasil akurasi yang sama juga pada setiap jumlah komponen PCA yaitu 70.54%. Pengujian dengan perbandingan sebesar 80%:20% memiliki akurasi tertinggi pada jumlah komponen PCA dari n=7 hingga n=10 yaitu 72.64%. Pengujian dengan perbandingan sebesar 83%:17%, didapatkan hasil akurasi tertinggi pada jumlah komponen PCA n=8 hingga n=10 yaitu 75.00%. Pengujian dengan perbandingan sebesar 85%:15%, didapatkan hasil akurasi tertinggi pada jumlah komponen PCA n=8 hingga n=10 yaitu 77.96%. Pengujian dengan perbandingan sebesar 90%:10%, didapatkan hasil akurasi tertinggi pada jumlah komponen PCA n=8 hingga n=10 yaitu 78.40%.

Hasil keseluruhan percobaan berdasarkan Tabel 3 dan Tabel 4 menunjukkan bahwa akurasi tertinggi didapatkan pada perbandingan *data training* dan *data testing* sebesar 90%:10% dengan komponen PCA n=8 hingga n=10 yaitu 78.40%, sehingga untuk mengurangi waktu komputasi, akurasi terbaik tepat pada jumlah komponen PCA n=8.

4. Pembahasan

Perbandingan performa klasifikasi model menggunakan algoritma decision tree dengan jumlah persentase *data training* dan *data testing* yang berbeda dan dengan penerapan PCA cukup beragam.

Pada penelitian sebelumnya oleh [4] untuk klasifikasi penyakit liver menggunakan algoritma decision tree dengan jumlah *data training* dan *data testing* sebesar 433:150 data, didapatkan hasil akurasi sebesar 72.67%. Pada penelitian ini, dengan jumlah perbandingan *data training* dan *data testing* yang hampir setara dengan penelitian sebelumnya yaitu sebesar 437:146 data (75%:25%), dengan penerapan menggunakan metode reduksi dimensi PCA dan algoritma Decision Tree, didapatkan akurasi sebesar 70.54%. Pada penelitian ini, dengan parameter yang sama dan di gunakan pendekatan reduksi dimensi PCA, menyebabkan akurasi menurun. Hal ini menerangkan bahwa penggunaan PCA tidak selalu bisa menaikkan akurasi, tetapi bisa juga menurunkan akurasi. Pernyataan ini dikuatkan oleh penelitian yang dilakukan oleh [19] yang menyatakan bahwa hasil penelitian yang dilakukan untuk klasifikasi kualitas air dengan menerapkan metode reduksi dimensi PCA dapat menurunkan tingkat performa metode klasifikasi kNN dan Logstic Regression, dimana dengan menggunakan seluruh fitur

tanpa PCA dengan algoritma kNN mendapatkan hasil terbaik yaitu akurasi 90.8% pada jumlah k=9.

Upaya untuk meningkatkan hasil akurasi setelah penggunaan metode reduksi dimensi PCA dengan cara mencoba persentase *data training* dan *data testing* yang berbeda. Pada penelitian ini dengan menggunakan jumlah persentase *data training* yang lebih banyak, dapat meningkatkan performa akurasi, sebagaimana hasil yang dapat dilihat pada Tabel 4. Tabel 4 menunjukkan hasil penggunaan PCA dengan jumlah *data training* terbesar yaitu 90% menghasilkan akurasi terbaik yaitu 78.40% pada jumlah komponen n=8. Penggunaan *data training* yang tinggi merupakan salah satu cara untuk meningkatkan akurasi, sebagaimana hasil penelitian terdahulu oleh [20] dalam implementasi algoritma Support Vector Machine (SVM) dalam menyelesaikan penyakit gejala demam dengan *data training* dan *data testing* sebesar 90%:10% menghasilkan akurasi terbaik yaitu 99.23% dibandingkan dengan jumlah pembagian 80%:20% dan 50%:50%. Penelitian sebelumnya juga dilakukan oleh [21] menghasilkan nilai akurasi degan menggunakan algoritma klasifikasi C4.5 yang terbesar pada percobaan *data training* dan *data testing* sebesar 90%:10% dibandingkan 80%:20% yaitu akurasi 85.34%.

Berdasarkan hasil pembahasan tersebut, untuk meningkatkan akurasi dalam klasifikasi penyakit liver dari penelitian-penelitian sebelumnya, didapatkan hasil akurasi terbaik yaitu sebesar 78.40% yang diperoleh dengan menggunakan algoritma decision tree yang sebelumnya telah di reduksi dimensi menggunakan PCA dengan jumlah komponen PCA n=8 dan persentase *data training* dan *data testing* sebesar 90%:10%.

5. Kesimpulan

Liver adalah organ penting dan terbesar manusia**.** Penyakit liver merupakan salah satu penyakit yang sulit dideteksi dan menjadi penyumbang kematian karena dianggap sebagai pembunuh diam-diam tanpa gejala. Penyakit liver dapat dideteksi berdasarkan kelainan jumlah kandungan-kandungan dalam tubuh manusia. Data pasien liver berdasarkan kandungan zat dalam tubuh dapat dijadikan data untuk menentukan apakah seseorang menderita penyakit liver atau tidak, dengan melakukan proses klasifikasi.

Proses klasifikasi penderita penyakit liver dapat dilakukan dengan metode Decision Tree serta menggunakan metode Principal Component Analysis (PCA) sebagai langkah preprocessing untuk mengurangi dimensi data, dan memilih jumlah persentase *data training* dan *data testing* yang tepat, agar hasil akurasi tinggi.

Hasil penelitian secara keseluruan yaitu tingkat performa dari metode Decision Tree menurun saat menerapkan metode reduksi dimensi PCA dengan *data training* dan testing sebesar 73%:27% dengan akurasi 69.62%, sedangkan jika *data training* dan *data testing* di ubah persentasenya menjadi 90%:10%, akan menghasilkan akurasi tertinggi yaitu 78.40% pada jumlah komponen PCA n=8.

Referensi

[1] C. Y. Gobel, "Sistem Pakar Penyakit Liver Menggunakan K- Nearest Neighbors Algoritm Berbasis Website," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 152–159, 2018, doi: 10.33096/ilkom.v10i2.296.152-159.

[2] N. T. Rahman, "Analisa Algoritma Decision Tree Dan Naïve Bayes Pada Pasien Penyakit Liver," *J. Fasilkom*, vol. 10, no. 2, pp. 144–151, 2020, doi: 10.37859/jf.v10i2.2087.

[3] B. L. Trust, "Liver disease is now the biggest cause of death in those aged between 35-49 years old, new report reveals," *British Liver Trust*, 2019. [Online]. Available: https://britishlivertrust.org.uk/liver-disease-is-now-the-biggest-cause-of-death-in-those-aged-between-35-49-years-old-new-report-reveals/. [Accessed: 11-Dec-2022].

[4] I. Setiawati, A. P. Wibowo, and A. Hermawan, "Implementasi Decision Tree Untuk Mendiagnosis Penyakit Liver," *J. Inf. Syst. Manag.*, vol. 1, no. 1, pp. 13–17, 2019.

[5] D.- Restiani, "Kombinasi Algoritma Cart Dan Ripper Untuk Mendiagnosis Penyakit Liver Berbasis Correlation Based Feature Selection," *J. Tek. Inform.*, vol. 11, no. 1, pp. 31–36, 2018, doi:

10.15408/jti.v11i1.6660.

[6] E. Muningsih, "Kombinasi Metode K-Means Dan Decision Tree Dengan Perbandingan Kriteria Dan Split Data," *J. Teknoinfo*, vol. 16, no. 1, p. 113, 2022, doi: 10.33365/jti.v16i1.1561.

[7] M. Windarti, "Perbandingan Kinerja Algoritma Naïve Bayes Dan Bayesian Network Dalam Klasifikasi Masa Studi Mahasiswa," *Pros. Semin. Nas. Apl. Sains Teknol.*, no. September, pp. 249–261, 2018.

[8] "Indian Liver Patient Dataset," *UCI Machine Learning Repository*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset).

[9] N. P. A. Widiari, I. M. A. D. Suarjaya, and D. P. Githa, "Teknik Data Cleaning Menggunakan Snowflake untuk Studi Kasus Objek Pariwisata di Bali," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 8, no. 2, p. 137, 2020, doi: 10.24843/jim.2020.v08.i02.p07.

[10] S. Raysyah, V. Arinal, and D. I. Mulyana, "Klasifikasi Tingkat Kematangan Buah Kopi Berdasarkan Deteksi Warna Menggunakan Metode Knn Dan Pca," *JSiI (Jurnal Sist. Informasi)*, vol. 8, no. 2, pp. 88–95, 2021, doi: 10.30656/jsii.v8i2.3638.

[11] A. Ilmaniati and B. E. Putro, "Analisis komponen utama faktor-faktor pendahulu (antecendents) berbagi pengetahuan pada usaha mikro, kecil, dan menengah (UMKM) di Indonesia," *J. Teknol.*, vol. 11, no. 1, pp. 67–78, 2019.

[12] R. A. Anggraini, G. Widagdo, A. S. Budi, and M. Qomaruddin, "Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Naïve Bayes," *J. Sist. dan Teknol. Inf.*, vol. 7, no. 1, p. 47, 2019, doi: 10.26418/justin.v7i1.30211.

[13] W. Musu, A. Ibrahim, and Heriadi, "Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5," in *Seminar Sistem Informasi dan Teknologi Informasi (SISITI)*, 2021, pp. 186–195.

[14] Y. Irawan, "Penerapan Algoritma Decision Tree C4.5 Untuk Memprediksi Kelayakan Calon Pendonor Melakukan Donor Darah Dengan Klasifikasi Data Mining," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 2, no. 4, pp. 181–189, 2021, doi: 10.35746/jtim.v2i4.75.

[15] S. Bahri, A. Lubis, U. Pembangunan, and P. Budi, "Metode Klasifikasi Decision Tree Untuk Memprediksi Juara English Premier League," *Sintaksis*, vol. 2, no. 04, pp. 63–70, 2020.

[16] M. H. Dunham, *Data Mining Introductory and Advanced Topics*. New Jersey: Prentice Hall, 2003.

[17] I. Sutoyo, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *J. PILAR Nusa Mandiri*, vol. 14, no. 2, pp. 217–224, 2018, doi: 10.35329/jiik.v7i2.203.

[18] D. A. H. D. Larasati and T. Sutrisno, "Tourism Site Recommendation in Jakarta Using Decision Tree Method Based on Web Review," *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.3268964.

[19] B. N. Azmi, A. Hermawan, and D. Avianto, "Analisis Pengaruh PCA Pada Klasifikasi Kualitas Air Menggunakan Algoritma K-Nearest Neighbor dan Logistic Regression," *JUSTINDO (Jurnal Sist. dan Teknol. Informasi)*, vol. 7, no. 2, pp. 94–103, 2022.

[20] N. I. Fadilah, B. Rahayudi, and M. T. Furqon, "Implementasi Algoritme Support Vector Machine (SVM) Untuk Klasifikasi Penyakit Dengan Gejala Demam," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, pp. 5619–5625, 2018.

[21] I. Alfarobi, T. A. Tutupoly, and A. Suryanto, "Komparasi Algoritma C4.5, Naive Bayes, Dan Random Forest Untuk Klasifikasi Data Kelulusan Mahasiswa Jakarta," *BSI Repos.*, 2017.