BAB IV
HASIL PENELITIAN DAN PEMBAHASAN

4.1 Hasil Penelitian

Berdasarkan metodologi yang telah dirancang untuk membandingkan tingkat akurasi tertinggi dalam memprediksi kasus penyakit *liver* dengan menggunakan algoritma *random forest,* C 5.0 dan *support vector machine.* maka hasil yang didapatkan adalah sebagai berikut :

4.1.1 *Data Collection*

Data yang diperoleh dari *kaggle* berjumlah 30.692 data penyakit *liver* dalam bentuk *comma separated values* (csv). Adapun data *sample* sebagai berikut:

	Age of the patient	Gender of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result
0	65	0	0	0	187	16	18	6	3	0	Yes
1	62	1	10	5	699	64	100	7	3	0	Yes
2	62	1	7	4	490	60	68	7	3	0	Yes
3	58	1	1	0	182	14	20	6	3	1	Yes
4	72	1	3	2	195	27	59	7	2	0	Yes

Gambar 4.1 Sampel Data

4.1.2 *Expolatory Data Analysis (EDA)*

a) Data Understanding

Dari hasil pemeriksaan, ditemukan adanya tipe data numerik dan kategori, serta beberapa kolom memiliki data kosong yang perlu ditangani.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30691 entries, 0 to 30690
Data columns (total 11 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Age of the patient                    30689 non-null  float64
 1   Gender of the patient                 29789 non-null  object
 2   Total Bilirubin                       30043 non-null  float64
 3   Direct Bilirubin                      30130 non-null  float64
 4    Alkphos Alkaline Phosphotase         29895 non-null  float64
 5    Sgpt Alamine Aminotransferase        30153 non-null  float64
 6   Sgot Aspartate Aminotransferase       30229 non-null  float64
 7   Total Protiens                        30228 non-null  float64
 8    ALB Albumin                          30197 non-null  float64
 9   A/G Ratio Albumin and Globulin Ratio  30132 non-null  float64
 10  Result                                30691 non-null  int64
dtypes: float64(9), int64(1), object(1)
memory usage: 2.6+ MB
```

Gambar 4.2 Informasi Data

b) *Duplicate*

Berikut hasil setelah dilakukan pemeriksaan, diketahui bahwa dataset mengandung 11323 data duplikat.

```
# Mengecek jumlah baris sebelum menghapus duplikat
print("Jumlah baris sebelum menghapus duplikat:", data.shape[0])

# Menghapus baris duplikat
data = data.drop_duplicates()

# Mengecek jumlah baris setelah menghapus duplikat
print("Jumlah baris setelah menghapus duplikat:", data.shape[0])
```

```
Jumlah baris sebelum menghapus duplikat: 30691
Jumlah baris setelah menghapus duplikat: 19368
```

Gambar 4.3 *Duplicate Data*

c) *Missing Value*

Setelah dilakukan pemeriksaan *missing value* diketahui bahwa dataset memiliki 5425 *missing values* yang kemudian dilakukan penghapusan. Hasil pengecekan pada setiap variabel adalah:

```
Age of the patient                          2
Gender of the patient                     902
Total Bilirubin                           648
Direct Bilirubin                          561
 Alkphos Alkaline Phosphotase             796
 Sgpt Alamine Aminotransferase            538
Sgot Aspartate Aminotransferase           462
Total Protiens                            463
 ALB Albumin                              494
A/G Ratio Albumin and Globulin Ratio      559
Result                                      0
dtype: int64
Total missing values in the dataset: 5425
```

Gambar 4.4 *Missing Value*

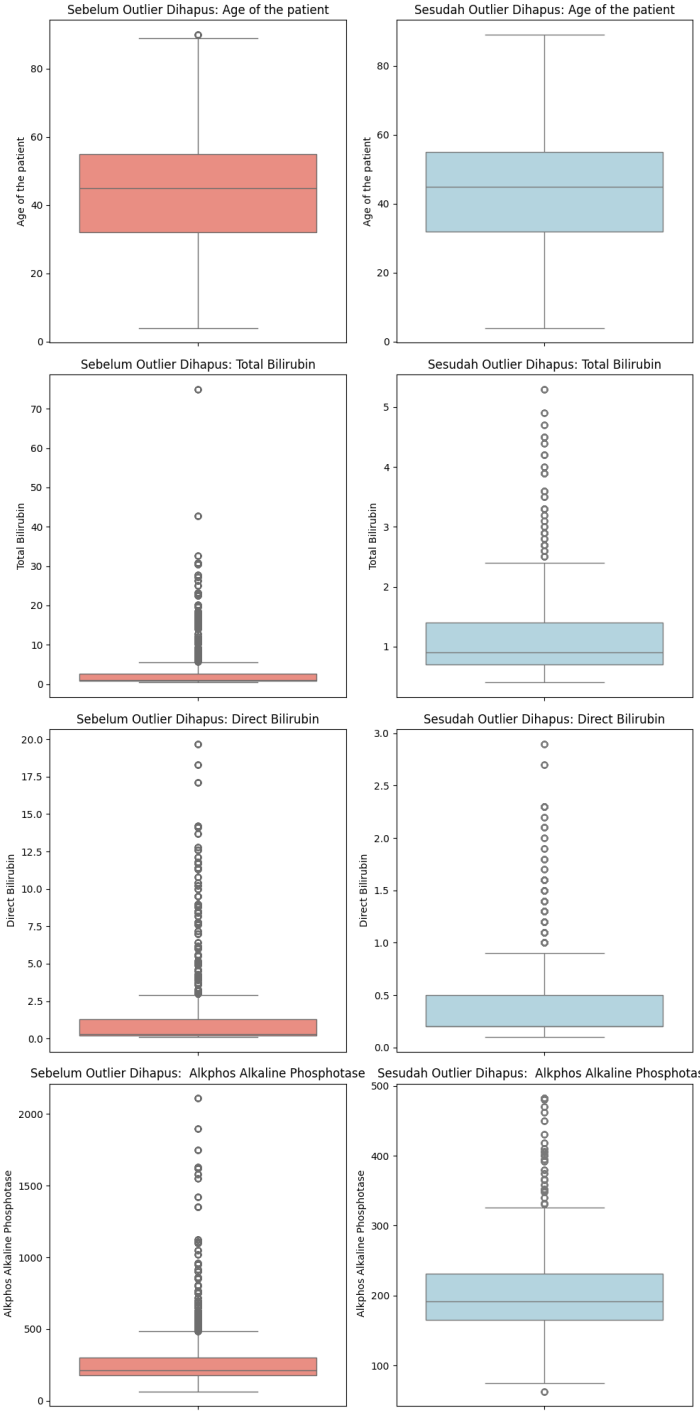
Output:

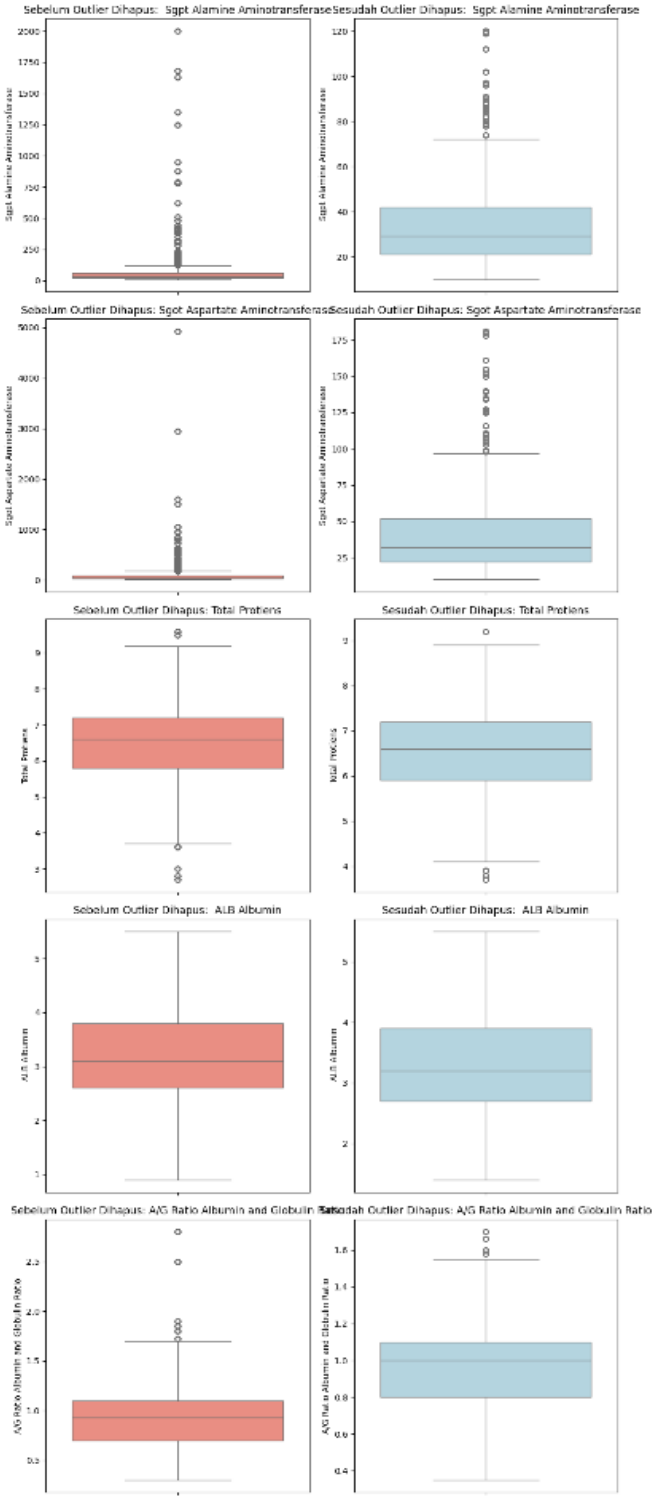
```
Jumlah data sebelum penghapusan missing value: 19368
Jumlah data setelah penghapusan missing value: 16389
```

Gambar 4.5 *Output Missing Value*

d) *Outlier*

Mengecek kemungkinan adanya *Outlier* dan memvisualisasikannya menggunakan boxplot, hasilnya adalah sebagai berikut:


Sebelum Outlier Dihapus: Age of the patient
Age of the patient
Sesudah Outlier Dihapus: Age of the patient
Age of the patient
Sebelum Outlier Dihapus: Total Bilirubin
Total Bilirubin
Sesudah Outlier Dihapus: Total Bilirubin
Total Bilirubin
Sebelum Outlier Dihapus: Direct Bilirubin
Direct Bilirubin
Sesudah Outlier Dihapus: Direct Bilirubin
Direct Bilirubin
Sebelum Outlier Dihapus: Alkphos Alkaline Phosphotase
Alkphos Alkaline Phosphotase
Sesudah Outlier Dihapus: Alkphos Alkaline Phosphotase
Alkphos Alkaline Phosphotase



Gambar 4.6 *Outlier*

Dalam memprediksi kasus penyakit *liver* ini *outlier* dapat memengaruhi performa model secara negative dan memungkinkan adanya *overfitting*, sehingga diputuskan untuk menghapusnya agar model dapat mengenali pola data secara lebih optimal. Setelah dilakukan pemeriksaan, dari 16389 data ditemukan *Outlier* berjumlah 5241 data yang kemudian dilakukan pengahapusan dan tersisa 11148 data.

```
Kolom: Age of the patient
Ambang Batas Atas (Maksimum): 89.5
Ambang Batas Bawah (Minimum): -2.5
Jumlah Outlier Melebihi Maksimum: 41
Jumlah Outlier Kurang dari Minimum: 0
--------------------------------------------------
Kolom: Total Bilirubin
Ambang Batas Atas (Maksimum): 5.550000000000001
Ambang Batas Bawah (Minimum): -2.05
Jumlah Outlier Melebihi Maksimum: 2396
Jumlah Outlier Kurang dari Minimum: 0
--------------------------------------------------
Kolom: Direct Bilirubin
Ambang Batas Atas (Maksimum): 2.95
Ambang Batas Bawah (Minimum): -1.4500000000000002
Jumlah Outlier Melebihi Maksimum: 2341
Jumlah Outlier Kurang dari Minimum: 0
--------------------------------------------------
Kolom:  Alkphos Alkaline Phosphotase
Ambang Batas Atas (Maksimum): 482.5
Ambang Batas Bawah (Minimum): -9.5
Jumlah Outlier Melebihi Maksimum: 1887
Jumlah Outlier Kurang dari Minimum: 0
--------------------------------------------------
Kolom:  Sgpt Alamine Aminotransferase
Ambang Batas Atas (Maksimum): 120.5
Ambang Batas Bawah (Minimum): -35.5
Jumlah Outlier Melebihi Maksimum: 1940
Jumlah Outlier Kurang dari Minimum: 0
--------------------------------------------------
Kolom: Sgot Aspartate Aminotransferase
Ambang Batas Atas (Maksimum): 182.5
Ambang Batas Bawah (Minimum): -69.5
Jumlah Outlier Melebihi Maksimum: 1924
Jumlah Outlier Kurang dari Minimum: 0
--------------------------------------------------
Kolom: Total Protiens
Ambang Batas Atas (Maksimum): 9.3
Ambang Batas Bawah (Minimum): 3.6999999999999993
Jumlah Outlier Melebihi Maksimum: 55
Jumlah Outlier Kurang dari Minimum: 176
--------------------------------------------------
Kolom:  ALB Albumin
Ambang Batas Atas (Maksimum): 5.6
Ambang Batas Bawah (Minimum): 0.8000000000000005
Jumlah Outlier Melebihi Maksimum: 0
Jumlah Outlier Kurang dari Minimum: 0
--------------------------------------------------
Kolom: A/G Ratio Albumin and Globulin Ratio
Ambang Batas Atas (Maksimum): 1.7000000000000002
Ambang Batas Bawah (Minimum): 0.09999999999999976
Jumlah Outlier Melebihi Maksimum: 297
Jumlah Outlier Kurang dari Minimum: 0
--------------------------------------------------
Kolom: Result
Ambang Batas Atas (Maksimum): 3.5
Ambang Batas Bawah (Minimum): -0.5
Jumlah Outlier Melebihi Maksimum: 0
Jumlah Outlier Kurang dari Minimum: 0
--------------------------------------------------
```

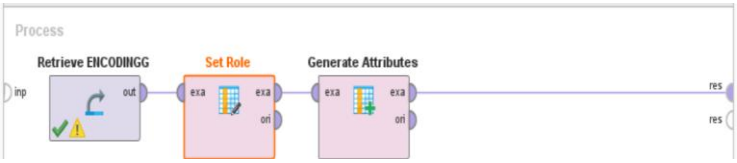
Gambar 4.7 *Upper lower*

4.1.3 *Data Prepocessing*

a) Data Encoding

Tahap ini mengubah atau mentransformasi data pada setiap fitur menjadi angka sehingga data dapat dilakukan proses training. Yaitu:

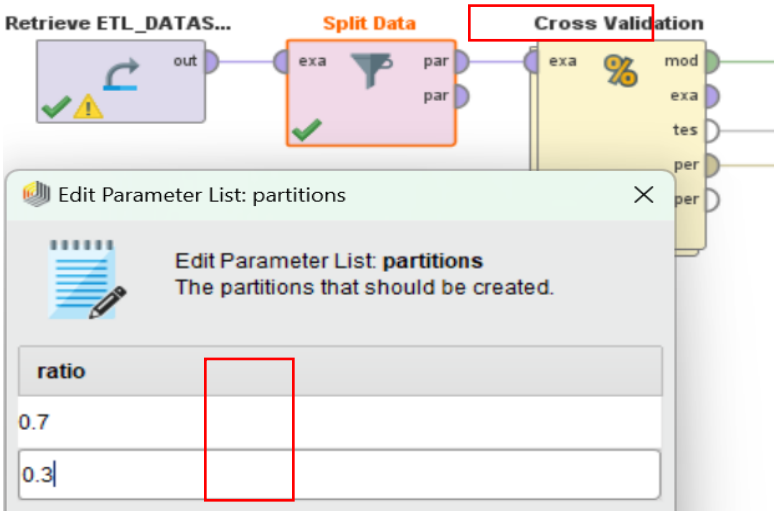
1. Kolom Gender, fitur Female akan diubah menjadi 0 dan fitur Male akan diubah menjadi 1.
2. Kolom Result, fitur 1 akan diubah menjadi Yes dan fitur 2 akan diubah menjadi No.



Gambar 4.8 Tahapan Data Encoding

b) Data Splitting

Berikut adalah proses *splitting data* di rapidminer:



Gambar 4.9 *Data splitting*

4.1.4 *Model Validation*

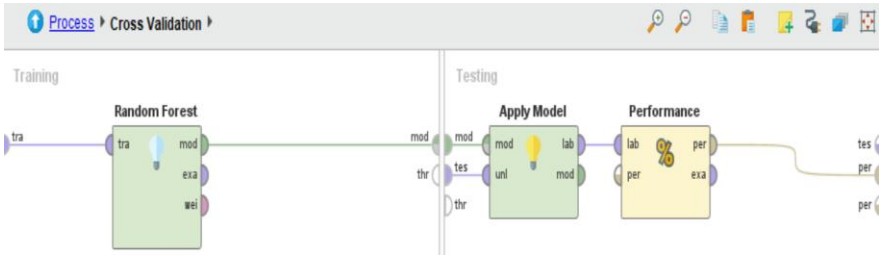
a. Confusion Matrix

Confusion matrix memberikan gambaran rinci mengenai hasil prediksi model dibandingkan dengan data sebenarnya. Matriks ini menunjukkan jumlah prediksi yang benar maupun salah untuk setiap kelas, sehingga memudahkan analisis performa model. Berikut tampilan hasil klasifikasi algoritma melalui *confusion matrix*:

accuracy: 96.78% +/- 1.35% (micro average: 96.78%)

	true Yes	true No	class precision
pred. Yes	4723	165	96.62%
pred. No	86	2830	97.05%
class recall	98.21%	94.49%	

Gambar 4.10 Hasil Klasifikasi *Random Forest*

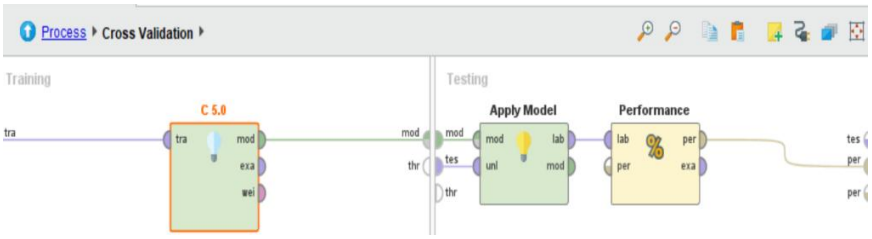


Gambar 4.11 Model RapidMiner *Random Forest*

accuracy: 65.26% +/- 0.72% (micro average: 65.26%)

	true Yes	true No	class precision
pred. Yes	4809	2711	63.95%
pred. No	0	284	100.00%
class recall	100.00%	9.48%	

Gambar 4.12 Hasil Klasifikasi C 5.0

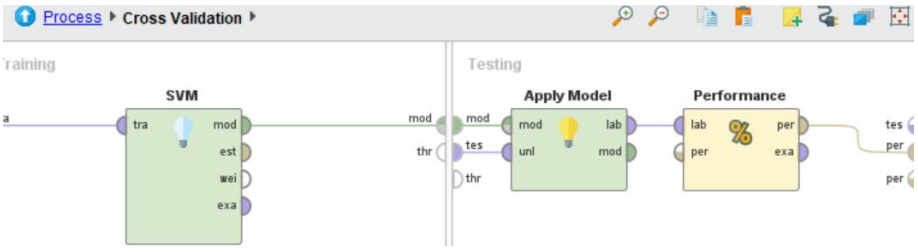


Gambar 4.13 Model RapidMiner C 5.0

accuracy: 73.44% +/- 1.28% (micro average: 73.44%)

	true Yes	true No	class precision
pred. Yes	4146	1410	74.62%
pred. No	663	1585	70.51%
class recall	86.21%	52.92%	

Gambar 4.14 Hasil Klasifikasi *Support Vector Machine*

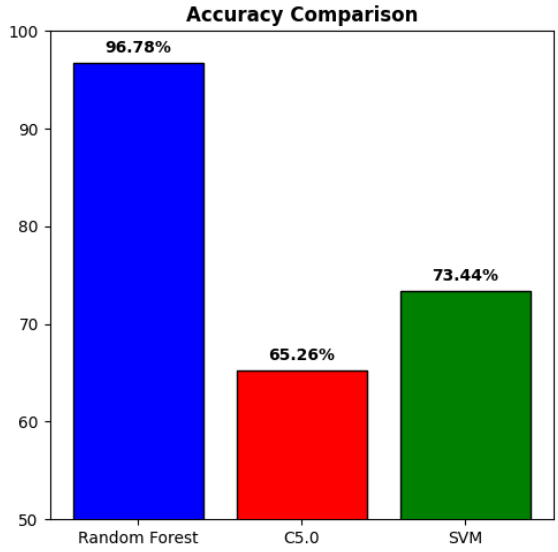


Gambar 4.15 Model RapildMiner *Support Vector Machine*

4.1.5 *Model Evaluation*

1. Akurasi

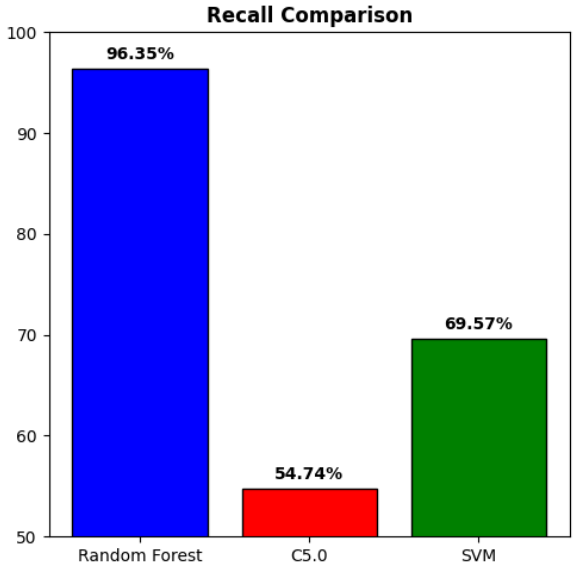
   Berikut adalah hasil perbandingan akurasi algoritma *random forest, C5.0, dan support vector machine (SVM*) dalam memprediksi penyakit liver.



Gambar 4.16 Akurasi

2. *Recall*

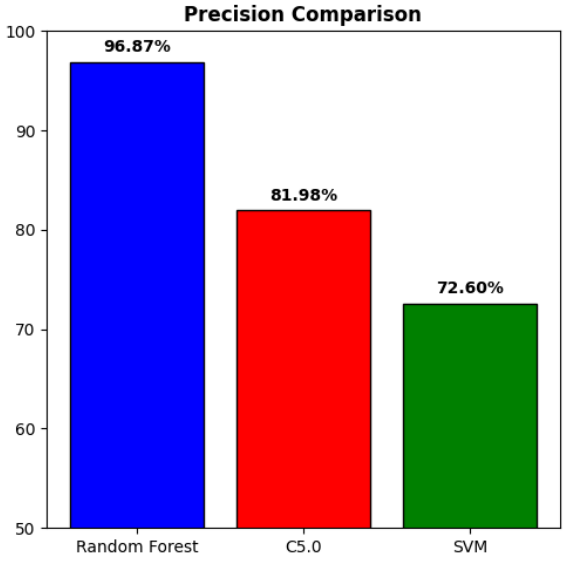
   Berikut adalah hasil perbandingan *recall* algoritma *random forest, C5.0, dan support vector machine* (SVM) dalam memprediksi penyakit liver.



Gambar 4.17 *Recall*

3. *Precision*

   Berikut adalah hasil perbandingan *precision* algoritma *random forest, C5.0, dan support vector machine* (SVM) dalam memprediksi penyakit liver.



Gambar 4.18 *Precision*

BAB V
KESIMPULAN

5.1 Kesimpulan

Berdasarkan hasil evaluasi performa model, algoritma Random Forest menunjukkan kinerja terbaik dengan akurasi sebesar 96.78%, recall sebesar 96.35%, dan precision sebesar 96.87%. Hal ini menunjukkan bahwa Random Forest mampu memberikan hasil prediksi yang sangat akurat dan efektif dalam mendeteksi kasus positif dengan tingkat kesalahan yang sangat rendah. Di sisi lain, algoritma C5.0 mencatat precision yang cukup tinggi sebesar 81.98%, tetapi memiliki akurasi (65.26%) dan recall (54.74%) yang rendah, sehingga kurang optimal dalam mendeteksi kasus positif secara konsisten. Sementara itu, algoritma Support Vector Machine (SVM) menunjukkan akurasi sebesar 73.44%, recall sebesar 69.57%, dan precision sebesar 72.60%, yang menunjukkan performa yang lebih baik dibandingkan C5.0 tetapi masih di bawah Random Forest. Secara keseluruhan, algoritma Random Forest menjadi pilihan terbaik dalam memprediksi risiko penyakit liver berdasarkan hasil evaluasi.

5.2 Saran

Saran-saran yang dapat penulis rekomendasikan pada penelitian lebih lanjut terkait kasus yang serupa adalah sebagai berikut:

a. Penelitian selanjutnya dapat menggunakan dataset yang lebih besar dan mencakup berbagai wilayah geografis serta kondisi kesehatan yang berbeda untuk meningkatkan kemampuan generalisasi model dalam memprediksi risiko penyakit liver.
b. Algoritma yang telah diimplementasikan dapat dikembangkan menjadi sistem deteksi dini berbasis aplikasi atau web, yang dapat

mempermudah dokter atau tenaga medis dalam mengidentifikasi risiko penyakit liver secara real-time.

c. Penelitian selanjutnya dapat memasukkan variabel prediktor tambahan, seperti riwayat penyakit keluarga, pola konsumsi alkohol, atau faktor genetik, untuk meningkatkan akurasi model dan memberikan hasil yang lebih komprehensif.
d. Penelitian di masa depan dapat memanfaatkan pendekatan *ensemble* atau *hybrid,* seperti menggabungkan *Random Forest* dengan *boosting* atau SVM dengan deep learning, untuk mendapatkan model yang lebih kuat dalam memprediksi risiko penyakit liver.