BAB III
METODOLOGI PENELITIAN

3.1 Subjek dan Objek Penelitian

Subjek penelitian ini merupakan algoritme klasifikasi dengan nilai akurasi terbaik pada penelitian sebelumnya, yaitu algoritme *Random Forest* dan *Support Vector Machine*s (*SVM*). Dari kedua algoritme tersebut akan dibandingkan kembali algoritme terbaik dengan nilai akurasi tertinggi dan waktu komputasi tercepat.

Objek penelitian ini adalah *dataset* mengenai penyakit hepatitis dengan variabel pada penelitian ini terdiri dari 19 variabel penjelas dan 1 variabel keputusan dengan label *Class* yaitu 0 (nol) dan 1 (satu). Label kelas pada data pasien penyakit hepatitis mengarah pada harapan hidup pasen penyakit hepatitis, apabila pasien penyakit hepatitis memiliki *variable* respon pada label *Class* DIE = 0 (nol) maka pasien penyakit hepatitis memiliki tingkat harapan hidup lebih tinggi dibandingkan pasien penyakit hepatitis yang memiliki variable kelas pada label *Class* LIVE = 1 (satu).

3.2 Teknik Pengambilan Data

Pengumpulan data yang diambil bersumber dari website UCI yang bersifat *public* ditulis oleh Gail Gong dari *Carnegie-Mellon University* melalui Bojan Cestnik dari *Institut Jozef Stefan*, Yugoslavia pada tanggal 1 November 1988 [71]. *Dataset* dapat diakses pada https://archive.ics.uci.edu/ml/datasets/Hepatitis.

3.3 Alat dan Bahan Penelitian

Alat :

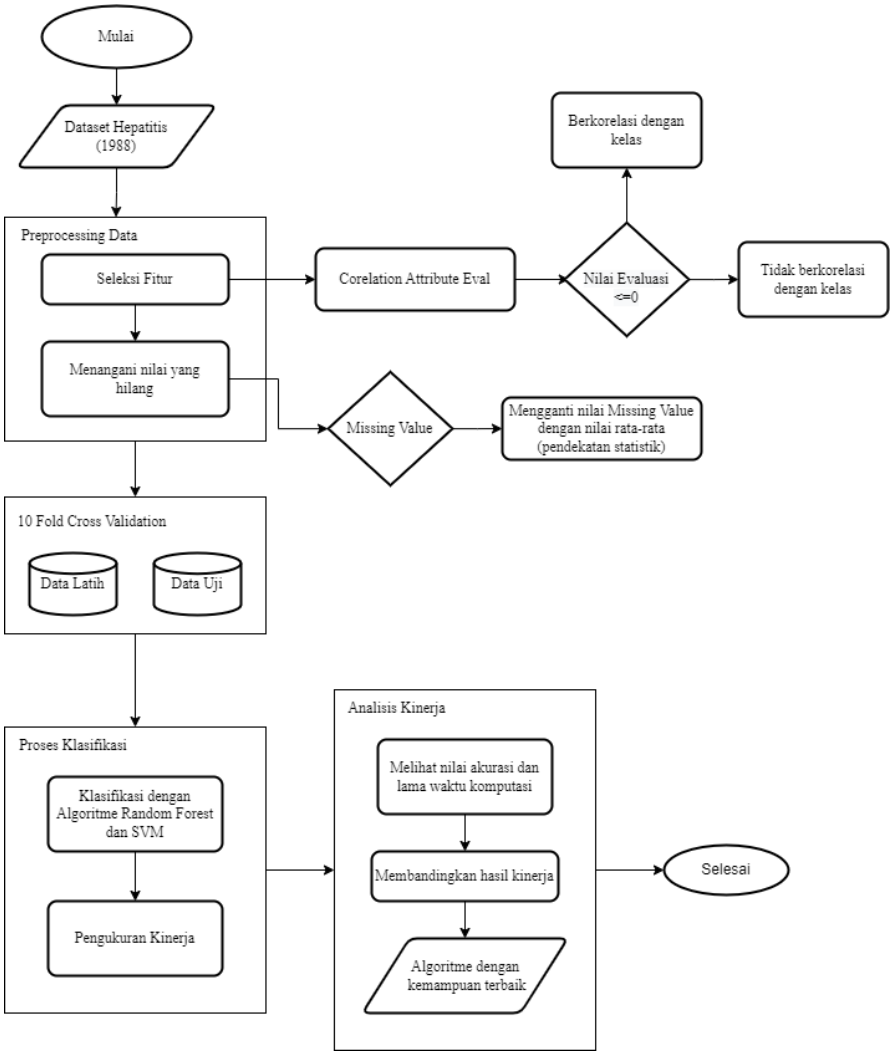
1. *Laptop ASUS AMD RYZEN3 3250U with Radeon Graphics* 2.60 GHz, RAM 8.00 GB.
2. *Rapidminer, Jupyter Notebook, WEKA.*

Bahan :

1. *Dataset* dari *website* UCI.

3.4 Diagram Alur Penelitian

Alur pada penelitian ini adalah langkah pertama yaitu menentukan topik penelitian, mengkaji permasalahan dalam bidang penelitian yang diambil, membuat tujuan, dan manfaat. Penelitian dilakukan dengan menggunakan data *public* yang diambil dari *website* UCI *Mechine Learning* berupa data pasien penyakit hepatitis dengan data berjumlah 155 *record*. Lalu melakukan proses klasifikasi dengan algoritme *Random Forest* dan *Support Vector Machine*s (*SVM*), penganalisisan algoritme dengan hasil kinerja terbaik sebagai penarikan kesimpulan sementara. Gambar 3.1 menunjukan diagram alur penelitian.



Gambar 3.1 Alur Penelitian

3.5 Dataset Hepatitis (1988)

Pengumpulan data yang diambil bersumber dari website UCI yang bersifat *public*. Adapun penjelasan mengenai *dataset* Hepatitis yang bersumber dari website UCI adalah sebagai berikut:

1. Data berjumlah 155 *record*.
2. *Dataset* berjumlah 19 atribut dan 1 atribut label yaitu *Class* [71].
3. *Dataset* hepatitis memiliki 167 nilai yang hilang dari data yang telah diperoleh *(missing value)* [71].

Pada Tabel 3.1 akan dijelaskan atribut yang terdapat pada dataset.

Tabel 3.1 Deskripsi Atribut *Dataset*

No.	Atribut	Domain Nama	Keterangan
1.	*Class*	*DIE*=0, *LIVE*=1	Merupakan kelas untuk menentukan harapan hidup a pasien hepatitis.
2.	*Age*	Angka numerik	Angka numerik umur pasien yang melakukan pemeriksaan.
3.	*Sex*	Laki-laki=1, Perempuan=2	Merupakan keterangan jenis kelamin pasien yang melakukan pemeriksaan.
4.	*Steroid*	*No*=1, *Yes*=2	Merupakan jenis obat yang dipakai untuk meningkatkan kemampuan fisik. Apakah pasien mendapatkan terapi steroid?
5.	*Antivirals*	*No*=1, *Yes*=2	Merupakan obat yang secara khusus digunakan untuk mengatasi penyakit yang disebabkan oleh infeksi virus. Apakah pasien mendapatkan terapi antiviral?
6.	*Fatigue*	*No*=1, *Yes*=2	Sebuah kondisi dimana tubuh merasa lelah, lesu dan tidak memiliki tenaga. Apakah pasien mengalami symstoms atau gejala kelelahan akut?
7.	*Malaise*	*No*=1, *Yes*=2	Dalam istilah medis yaitu perasaan tidak nyaman, lelah, dan kurang enak badan yang tidak diketahui penyebabnya. Apakah

No.	Atribut	Domain Nama	Keterangan
			pasien mengalami gejala malaise atau rasa tidak nyaman?
8.	*Anorexia*	*No*=1, *Yes*=2	Merupakan gangguan makan dimana ditandai dengan gejala muntah setiap makan.
9.	*Liver Big*	*No*=1, *Yes*=2	Apakah kondisi pasien dengan kondisi hati/liver membesar?
10.	*Liver Firm*	*No*=1, *Yes*=2	Apakah kondisi pasien dengan kondisi hati/liver membesar?
11.	*Spleen Palpable*	*No*=1, *Yes*=2	Suatu kondisi dimana limfa mengalami pembesaran, ukuran cenderung lebih besar dari normal.
12.	*Spiders*	*No*=1, *Yes*=2	Merupakan suatu gejala dimana pembuluh darah mengumpul dan menonjol pada permukaan kulit pasien.
13.	*Ascites*	*No*=1, *Yes*=2	Terjadi penumpukan cairan pada rongga antara selaput yang melapisi dinding perut dan ordan dalam tubuh.
14.	*Varices*	*No*=1, *Yes*=2	Kondisi dimana vena tidak bekerja dengan baik dikarenakan terjadi pembekakan pada vena.
15.	*Bilirubin*	Angka Numerik	Nilai kadar zat yang terbentuk secara normal dari proses penguraian sel darah merah di dalam tubuh.
16.	*Alkaline Phosphate*	Angka Numerik	Kadar Alkalin Phospate dalam liver atau hati yang bermanfaat untuk mengukur seberapa baik organ hati dan kantung empedu seseorang atau pasien.
17.	*Sgot*	Angka Numerik	Merupakan enzim yang ditemukan pada organ hati (liver), jantung, ginjal, hingga otak.
18.	*Albumin*	Angka Numerik	Kadar protein pada darah yang membentuk sebagian plasma darah.

No.	Atribut	Domain Nama	Keterangan
19.	*Protime*	Angka Numerik	Merupakan uji masa protrombhine (protein yang membantu proses pembekuan darah).
20.	*Histology*	*No*=1, *Yes*=2	Merupakan cabang ilmu biologi yang mempelajari tentang struktur sel dan jaringan secara detail menggunakan alat mikroskop. Apakah pasien dilakukan pemeriksaan sampel jaringan hati untuk memeriksa kondisi kesehatan pada hati?

Atribut pada *dataset* dijelaskan pada Tabel sampel record *dataset* Hepatitis. Sampel *record* dataset dapat dilihat pada Tabel 3.2.

Tabel 3.2 *Sampel Record* Dataset Hepatitis (1988-11-01)

Atribut	Dataset						
	1	2	3	4	5	…	155
Class	1	1	1	1	1	…	0
Age	30	50	78	31	34	…	43
Sex	2	1	1	1	1	…	1
Steroid	1	1	2	?	2	…	2
Antivirals	2	2	2	1	2	…	2
Fatigue	2	1	1	2	2	…	1
Malaise	2	2	2	2	2	…	2
Anorexia	2	2	2	2	2	…	2
Liver Big	1	1	2	2	2	…	2
Liver Firm	2	2	2	2	2	…	2
Spleen Palpable	2	2	2	2	2	…	1
Spiders	2	2	2	2	2	…	1
Ascites	2	2	2	2	2	…	1
Varices	2	2	2	2	2	…	2
Bilirubin	1	0.9	0.7	0.7	1	…	1.2
Alkaline Phosphate	85	135	96	46	?	…	100
Sgot	18	42	32	52	200	…	19
Albumin	4	3.5	4	4	4	…	3.1

Atribut	Dataset						
	1	2	3	4	5	…	155
Protime	?	?	?	80	?	…	42
Histology	1	1	1	1	1	…	2

3.6 *Preprocessing* Data

Preprocessing data pada penelitian ini dilakukan dengan tujuan untuk mengolah data agar siap digunakan. *Preprocessing Data* pada penelitian ini yaitu melakukan Seleksi Fitur dan *Handling Missing Value*.

3.6.1 *Feature Selection*

Hasil dari seleksi fitur adalah sebagai berikut:

1. Dataset Hepatitis berisi 167 nilai yang hilang atau kurang lebih sebesar 5,7% dari keseluruhan dataset [72] [73]. Pada penelitian ini target utama kelas dalam *dataset* yaitu atribut *Class* sesuai dengan penelitian [74].
2. Pada penelitian ini menggunakan metode *Correlation Atributte Eval* merupakan metode seleksi fitur pencarian *ranking* untuk mengevaluasi korelasi antara atribut dengan antribut yang dijadikan kelas utama [75]. Metode *Correlation Atributte Eval* menggunakn persamaan sebagai berikut [76]:

$$Merit = \frac{\overline{kavg(corr_{fc})}}{\sqrt{k + k(k-1)\overline{avg(corr_{ff})}}} \tag{3.1}$$

Keterangan :

Merit = korelasi antara set fitur dan kelas dependen,

K = jumlah fitur,

(avg(corrfc)) = rata-rata korelasi antar fitur dan kelas dependen,

avg(corrff) = rata-rata korelasi antara fitur.

Dengan menggunakan metode tersebut didapat atribut dengan peringkat tertinggi yaitu atribut nomor 13, 18, 12, 14, 7, 6, 19, 11, 3, 4, 8, 10. Sedangkan atribut dengan peringkat terendah yaitu atribut nomor

15, 20, 2, 16, 5, 17, 9. Total atribut yang dipakai sesuai metode ini adalah 12 atribut. Atribut dengan nilai ≤ 0 maka tidak memiliki korelasi dengan atribut label yaitu atribut *class*, sebaliknya apabila atribut dengan nilai > 0 maka memiliki korelasi dengan atribut label yaitu atribut *class*. Pada Gambar 3.2 menunjukkan hasil pengolahan dataset metode *Correlation Atributte Evaluation* menggunakan *software* WEKA.

```
Ranked attributes:
 0.4703  13 Ascites
 0.4658  18 Albumin
 0.392   12 Spiders
 0.363   14 Varises
 0.3376   7 Malaise
 0.309    6 Fatigue
 0.3072  19 Protime
 0.2349  11 Spleenpalpable
 0.1731   3 Sex
 0.1346   4 Steroid
 0.1321   8 Anorexia
 0.0587  10 Liverfirm
-0.0722   9 Liverbig
-0.0756  17 Sgot
-0.1302   5 Antivirals
-0.141   16 Alkphosphate
-0.2196   2 Age
-0.3379  20 Histology
-0.4505  15 Bilirubin
```

Gambar 3.2 Hasil *Feature Selection* Menggunakan WEKA

Total atribut pada *dataset* Hepatitis berjumlah 20 akan dipakai sesuai metode *Correlation Atributte Eval* yaitu berjumlah 12 atribut ditambah 1 atribut sebagai kelas utama yaitu *Class*.

3.6.2 *Handling Missing Value*

Dataset Hepatitis terdapat 12 atribut yang digunakan, ada 12 atribut *missing value* yaitu data atribut yang terdapat nilai "?" pada dataset. Atribut dengan *missing value* dapat dilihat pada Tabel 3.3.

Tabel 3.3 Atribut *Missing Value*

No. Atribut	**Atribut**	**Jumlah *Missing Value***
1.	*Class*	0
2.	*Sex*	0
3.	*Steroid*	1
4.	*Fatigue*	1
5.	*Malaise*	1
6.	*Anorexia*	1
7.	*Liver Firm*	11
8.	*Spleen Palpable*	5
9.	*Spiders*	5
10.	*Ascites*	5
11.	*Varices*	5
12.	*Albumin*	16
13.	*Protime*	68

Nilai yang hilang dapat mempengaruhi nilai akurasi. Penanganan *missing value* dilakukan dengan cara mengganti data pada atribut-atribut *missing value* menggunakan cara menghitung nilai pengganti *(imputation)* menggunakan nilai rata-rata dari masing-masing kelas yang mana adalah pendekatan statistik [77]. Persamaan untuk menghitung nilai rata-rata adalah sebagai berikut:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.2}$$

Keterangan :

xi = jumlah data ke i.

n = jumlah data.

Proses *handling missing value* pada atribut *Steroid, Fatigue, Malaise, Anorexia, Liver Firm, Spleen Palpable, Spiders, Ascites, Varices, Albumin,* dan *Protime* data dengan atribut yang nilainya "?" pada *dataset* akan dihitung nilai rata-rata dari masing-masing atribut. Tabel 3.4 merupakan hasil proses perhitungan nilai rata-rata pada atribut *missing value* adalah sebagai berikut :

Tabel 3.4 Hasil Perhitungan *Missing Value* Tiap Atribut

No. Atribut	Atribut	Hasil Perhitungan	Pembulatan
1.	*Steroid*	1.5064935064936	1
2.	*Fatigue*	1.35064935064935	1
3.	*Malaise*	1.6038961038961	2
4.	*Anorexia*	1.79220779220779	2
5.	*Liver Firm*	1.58333333333333	1
6.	*Spleen Palpable*	1.8	2
7.	*Spiders*	1.6	2
8.	*Ascites*	1.86666666666666	2
9.	*Varices*	1.88	2
10.	*Albumin*	3.81726618705036	3.8
11.	*Protime*	62.5632183908046	62

3.6.3 Dataset Baru

Pada Tabel 3.5 merupakan tabel yang berisi dataset yang telah dilakukan *preprocessing*.

Tabel 3.5 Dataset Baru Hasil *Preprosessing*

Atribut	Dataset						
	1	**2**	**3**	**4**	**5**	**…**	**155**
Class	2	2	2	2	2	…	1
Sex	2	1	1	1	1	…	1
Steroid	1	1	2	1	2	…	2
Fatigue	2	1	1	2	2	…	1
Malaise	2	2	2	2	2	…	2
Anorexia	2	2	2	2	2	…	2
Liver Firm	2	2	2	2	2	…	2
Spleen Palpable	2	2	2	2	2	…	1
Spiders	2	2	2	2	2	…	1
Ascites	2	2	2	2	2	…	1
Varices	2	2	2	2	2	…	2
Albumin	4	3.5	4	4	4	…	3.1
Protime	62,5632	62,5632	62,5632	80	62,5632	…	42

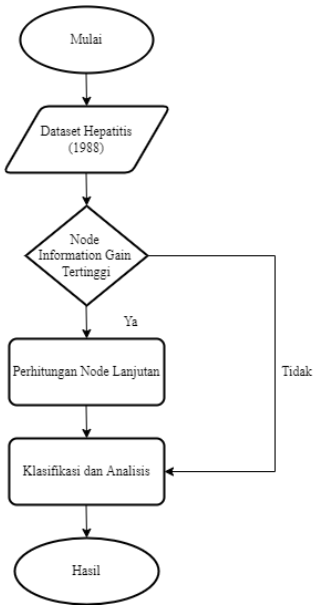
3.7 10-*Flod Cross Validation*

Pengujian pada penelitian ini menggunakan K-*Fold Cross Validation* dengan k=10. Pembagian data latih digunakan untuk membangun model dan data uji untuk menguji model yang telah dibangun.

3.8 Alur Algoritme

3.8.1. Alur *Random Forest*

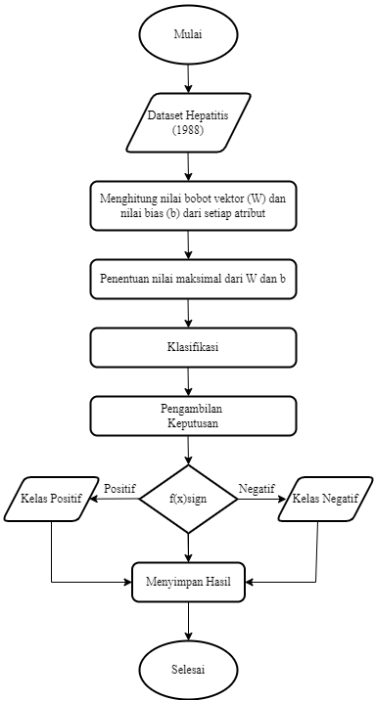
Random Forest merupakan metode berbasis klasifikasi yang memiliki beberapa tahapan dalam proses eksekusinya. Alur algoritme *Random Forest* dapat terlihat pada Gambar 3.4.



Gambar 3.3 Diagram Alur Algoritme *Random Forest*

3.8.2. Algoritme *Support Vector Machine*s (*SVM*)

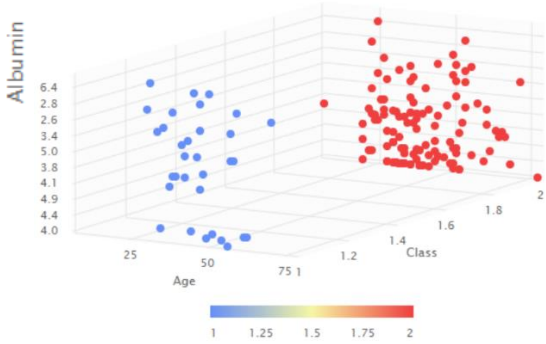
*Support Vector Machine*s (*SVM*) merupakan algoritme klasifikasi yang memiliki beberapa tahapan dalam proses eksekusinya. Alur algoritme *Support Vector Machine*s (*SVM*) dapat terlihat pada Gambar 3.5.



Gambar 3.4 Diagram Alur Algoritme *SVM*

3.9 Analisis Data

Analisis data pada penelitian ini akan diklasifikasikan dengan menggunakan 2 (dua) algoritme yaitu *Random Forest* dan *Support Vector Machine* (*SVM*). Hasil klasifikasi dari kedua algoritme tersebut akan dibandingkan dari segi nilai akurasi dan lama waktu komputasi sebagai penentu algoritme yang memiliki kemampuan paling baik. Gambar 3.6 merupakan *visualisasi* persebaran data pada *dataset* hepatitis, data berbentuk *non*-linier karena data tersebar tidak membentuk suatu garis linier.



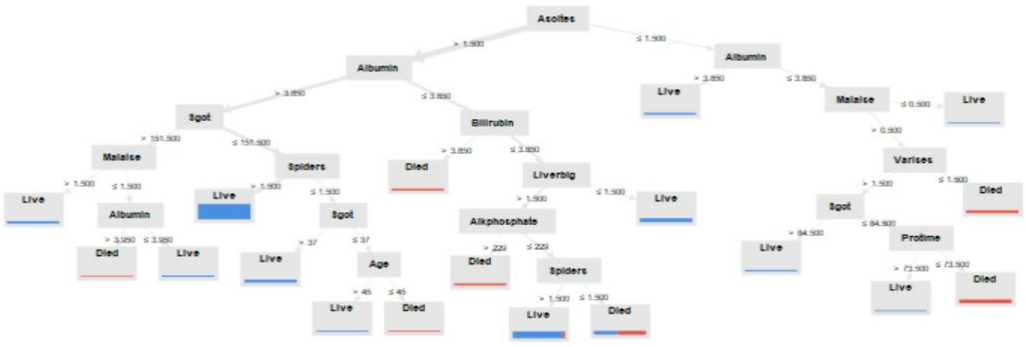
Gambar 3.5 Persebaran Dataset Hepatitis

Hasil *dataset* baru yang telah dilakukan *preprocessing* sesuai dengan Tabel 3.5 akan dilakukan eksperimen awal menggunakan *software* rapidminer dengan melakukan klasifikasi terhadap data tersebut menggunakan algoritme *Random Forest* dan *SVM*. Pembagian data latih dan uji menggunakan 10-*cross validation*. Algoritme *Random Forest* pada pengujian ini menggunakan *number of trees* = 10 dan dihasilkan *maximal depth* terbaik yang diujikan adalah 7 dari rentang 1 sampai dengan 10. Pencarian *maximal depth* dengan 10 kali percobaan dapat dilihat pada Tabel 3.6.

Tabel 3.6 Pencarian *Maximal Depth*

Percobaan	***Max Depth***	**Akurasi**
1	1	79.35%
2	2	79.35%
3	3	81.94%
4	4	82.58%
5	5	81.94%
6	6	84.52%
7	7	85.81%
8	8	85.16%
9	9	83.23%
10	10	82.58%

Tree yang dihasilkan pada pengujian dengan algoritme *Random Forest* dapat dilihat pada Gambar 3.6.



Gambar 3.6 Tree Pengujian Menggunakan Algoritme *Random Forest*

Hasil dari pengujian menggunakan algoritme *Random Forest* dapat dilihat pada Tabel 3.7.

Tabel 3.7 Hasil Klasifikasi dengan Algoritme *Random Forest*

	***True* Live**	***True* Died**	***Class Precission***
Pred. Live	114	13	89.76%
Pred. Died	9	19	67.86%
Class Recall	92.68%	59.38%	
Accuracy	85.81%		

Berdasarkan Tabel 3.7 hasil klasifikasi data sebenarnya yang *live* ketika dipresiksi juga *live* sebanyak 114, data sebenarnya yang *died* ketika dipresiksi *live* sebanyak 13, data sebenarnya yang *live* ketika dipresiksi *died* sebanyak 9, dan data sebenarnya yang di presiksi *died* sebanyak 19. Algoritme *Random Forest* dengan *class precission* sebesar 89.76% untuk prediksi *live* dan *class precission* prediksi *died* yaitu sebesar 67.86%. *Class recall* untuk *true live* sebesar 92.68% sedangkan *true died* sebesar 59.38%. Hasil klasifikasi Hasil klasifikasi algoritme ini menunjukan nilai akurasi sebesar 81.29%.

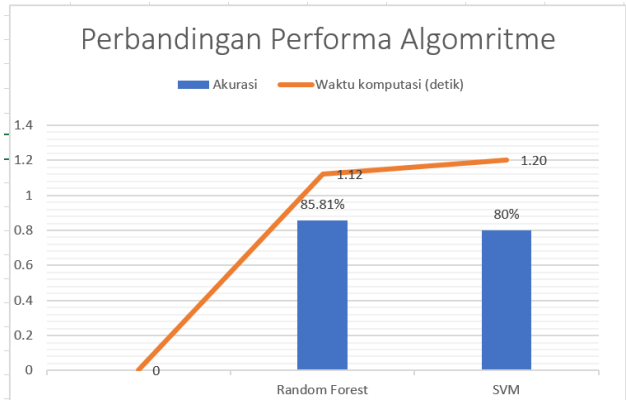
Algoritme *SVM* pada penelitian ini menggunakan *kernel Radial Basic Function* (RBF) karena fungsi *kernel* RBF digunakan klasifikasi data *non*-linier dan dikenal memiliki performa yang baik dengan parameter tertentu serta hasil dari pelatihan memiliki *error* yang kecil dibandingkan dengan *kernel* lainnya [78]. Hasil dari pengujian menggunakan algoritme *Random Forest* dapat dilihat pada Tabel 3.8.

Tabel 3.8 Hasil Klasifikasi dengan Algoritme *SVM*

	***True* Live**	***True* Died**	***Class Precission***
Pred. Live	117	25	82.39%
Pred. Died	6	7	53.85%
Class Recall	95.12%	21.88%	
Accuracy	80.00%		

Berdasarkan Tabel 3.8 hasil klasifikasi data sebenarnya yang *live* ketika dipresiksi juga *live* sebanyak 117, data sebenarnya yang *died* ketika dipresiksi *live* sebanyak 25, data sebenarnya yang *live* ketika dipresiksi *died* sebanyak 6, dan data sebenarnya yang di presiksi *died* sebanyak 7. Algoritme *SVM* dengan *class precission* sebesar 79.35% untuk prediksi *live* lebih unggul dari *class precission* prediksi *died* yaitu sebesar 0.00%. *Class recall* untuk *true live* mencapai nilai 100.00% sedangkan *true died* sebesar 0.00%. Hasil klasifikasi algoritme ini menunjukan nilai akurasi sebesar 79.35%.

Gambar 3.7 merupakan visualisasi hasil perbandingan algoritme *Random Forest* dan *SVM* berdasarkan akurasi dan waktu komputasi.



Gambar 3.7 Perbandingan Akurasi dan Waktu Komputasi

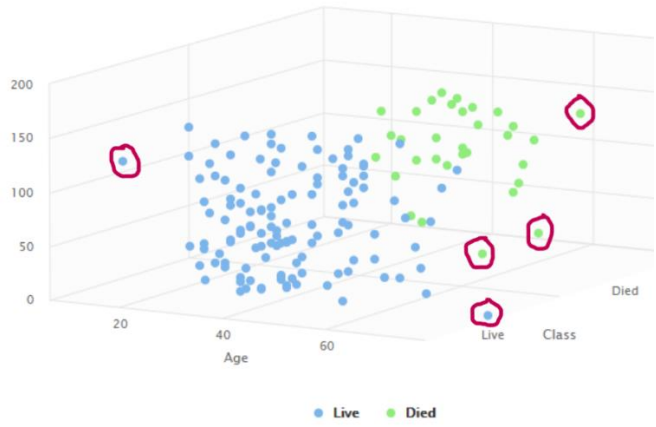
Berdasarkan Gambar 3.7 Algoritme *Random Forest* lebih unggul dibandingkan Algoritme *SVM*, dilihat dari perhitungan waktu komputasi yang dilakukan menggunakan *stopwatch* pada *smartphone* menghasilkan durasi lama waktu komputasi lebih rendah sehingga menghasilkan akurasi yang lebih tinggi. Hasil perbandingan klasifikasi algoritme *Random Forest* dan *SVM* dapat dilihat pada Tabel 3.9.

Tabel 3.9 Hasil Perbandingan Klasifikasi Algoritme *RF* dan *SVM*

Parameter	***RF***	***SVM***
Recall Kelas *Live*	92.68%	95.12%

Parameter	***RF***	***SVM***
Recall Kelas *Died*	59.38%	21.88%
Precission Prediksi *Live*	89.76%	82.39%
Precission Prediksi *Died*	67.86%	53.85%
Akurasi	85.81%	80.00%
Waktu Komputasi	01.12 detik	01.20 detik

Tabel 3.9 menunjukan hasil perbandingan klasifikasi kedua algoritme yang dibandingkan. Algoritme *Random Forest* memiliki nilai akurasi yang lebih baik dibandingkan *SVM*. Perbandingan dari segi lama waktu komputasi juga didapat *Random Forest* memiliki kemampuan lebih baik dibandingkan *SVM*, lama waktu komputasi menggunakan algoritme *Random Forest* lebih cepat dibandingkan dengan algoritme *SVM*. Dapat disimpulkan algoritme *Random Forest* memiliki kemampuan yang lebih unggul. Algoritme *Random Forest* dalam proses pengambilan keputusan dibantu dengan adanya *tree* atau pohon keputusan dan *max depth* terbaik sehingga dapat membuat nilai akurasi semakin tinggi. Sedangkan algoritme *SVM* dengan data bersifat *non*-linier data tersebar dan memungkinkan terdapat *outlier*, hal tersebut mengakibatkan hasil nilai akurasi rendah. *Outlier* adalah kasus data dengan karakteristik unik yang terlihat diluar kumpulan data pada suatu persebaran data [79]. Pengecekan *outlier* pada penelitian ini dilakukan menggunakan *software* rapidminer, hasil dari pengecekan *outlier* dapat dilihat pada Gambar 3.8.



Gambar 3.8 Pengecekan *Outlier*