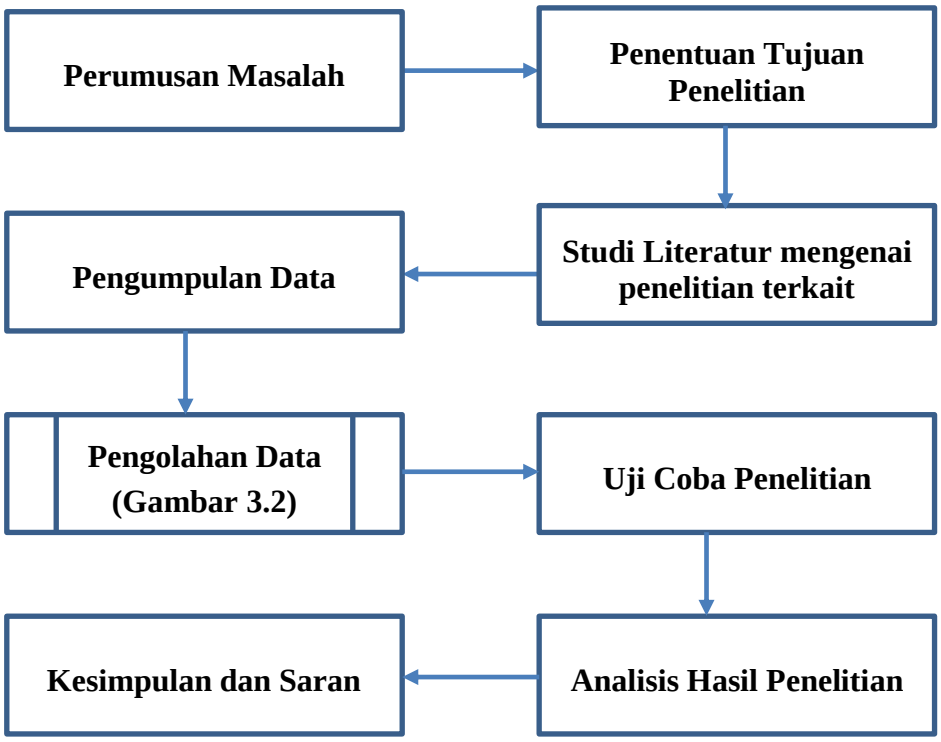
BAB III
METODOLOGI PENELITIAN

3.1 Desain Penelitian

Tahapan yang dilakukan dalam penelitian ini akan dibuat sebuah kerangka kerja sehingga penelitian akan sesuai dengan alur yang sudah direncanakan sebelumnya.

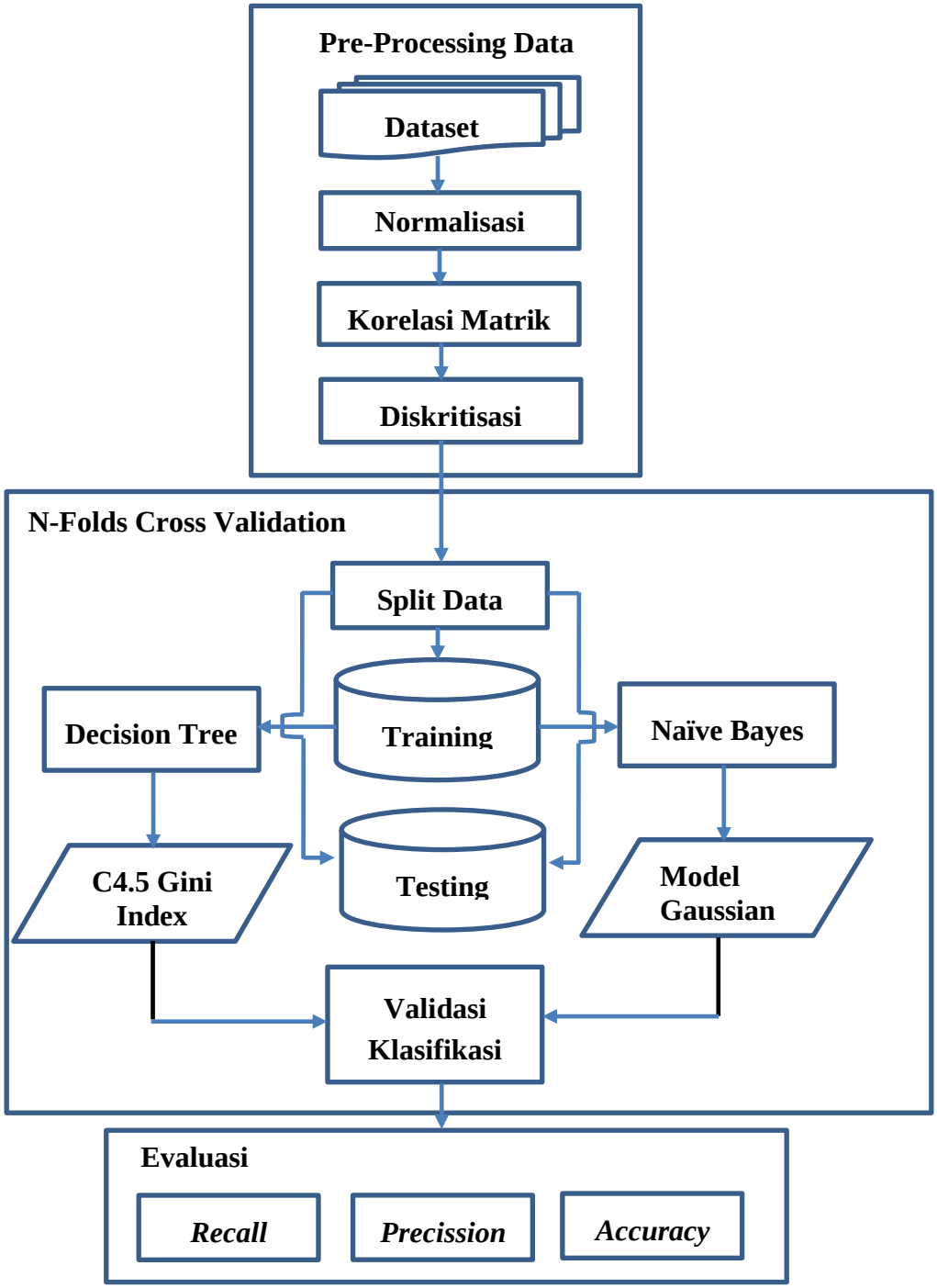
Proses pertama adalah menemukan sebuah masalah yang akan dijadikan topik pada penelitian ini. Kemudian masalah tersebut akan dikaji lebih dalam dengan mencari berbagai literature yang berkaitan dengan masalah yang akan diangkat.

Dalam penelitian ini peneliti menggunakan metode Algoritme Naïve Bayes Clasiffier dan Decision Tree dalam mendeteksi malware. Alur tahapan penelitian yang dibuat sebagai kerangka kerja untuk pendeteksian malware bisa dilihat pada Gambar dibawah ini :



Gambar 3.1 Kerangka kerja

Adapun proses pengolahan data secara detail ditunjukkan pada Gambar 3.2 seperti di bawah ini :



Gambar 3.2 Diagram Proses Pengolahan Data

Adapun prosesnya dimulai dengan tahap pertama yaitu pra-processing data yakni membaca dataset berupa data dengan format Comma Separated Values (CSV), lalu dataset tesebut di normalisasi dengan metode Z-Score dengan tujuan untuk mengurangi kesalahan pada proses pembacaan data, selanjutnya dilakukan proses menghilangkan atribut-atribut yang tidak kontributif dengan metode korelasi matrik. Proses terakhir pada tahapan ini yakni proses diskritisasi dalam dataset dengan metode binning yang dilakukan untuk menyesuaikan terhadap kemungkinan munculnya nilai kontinyu dalam karakteristik dataset yang kecil sehingga akan membawa pengaruh dalam proses klasifikasi.

Tahap kedua adalah tahap pemodelan data dengan menggunakan implementasi algoritma Decision Tree dan naïve bayes dengan teknik n-Folds Cross Validation. Teknik n-Fold Cross Validation ini adalah metode validasi silang. Metode ini membagi data menjadi dua bagian, yaitu data pelatihan (training) dan data pengujian (testing). Selanjutnya setelah data diuji dilakukan proses silang dimana data pegujian lantas dijadikan data pelatihan ataupun sebaliknya, data pelatihan sebelumnya dijadikan kini menjadi data pengujian. Data training yang sudah siap dilakukan proses pemodelan klasifikasi menggunakan algoritma decision tree model C4.5 Gini Index sedangkan untuk naïve bayes mneggunakan model Gaussian, setelah model-model klasifikasi tersebut didapatkan lalu diimplementasikan pada data testing. Tahap yang terakhir yakni proses evaluasi klasifikasi dengan melakukan pengukuran terhadap sensitifitas(recall), precision dan akurasi.

3.2 Spesifikasi Hardware dan Sofware

Eksperimen dilakukan pada Hardware Notebook Asus A409 FJ Intel Core ™ I7- 8565U CPU @ 1,80 GHz 1.99 GHz RAM 8 GB dan Software Windows 10 Home Single Language Include Java 8 Update 92 serta menggunakan Rapidminer Studio Educational 9.3.001 unntuk pengolahan dataset.

3.3 Dataset

Penelitian ini menggunakan dataset dengan tipe file CSV (Comma Separated Values) yang berextensi file excel. CSV (Comma Separated Values) merupakan suatu format data dalam basis data dimana setiap record dipisahkan dengan tanda koma (,) atau titik koma (;). Data yang digunakan adalah data sekunder dari dataset malware yang diambil dari website kaggle milik saravana (Saravana, 2018). Jumlah dataset malware yang digunakan peneliti yaitu 100.000 instance yang berisi data numerik (kontinyu) dengan 34 attribut seperti pada Tabel dibawah ini :

Tabel 3.1 Tipe Data Attribute Dataset Malware

No	Attribut	Type Data
1	millisecond	Numeric
2	classification	String
3	state	Numeric
4	usage_counter	Numeric
5	prio	Numeric
6	static_prio	Numeric
7	normal_prio	Numeric
8	policy	Numeric
9	vm_pgoff	Numeric
10	vm_truncate_count	Numeric
11	task_size	Numeric
12	cached_hole_size	Numeric
13	free_area_cache	Numeric
14	mm_users	Numeric
15	map_count	Numeric
16	hiwater_rss	Numeric
17	total_vm	Numeric
18	shared_vm	Numeric
19	exec_vm	Numeric
20	reserved_vm	Numeric
21	nr_ptes	Numeric
22	end_data	Numeric
23	last_interval	Numeric
24	nvcsw	Numeric
25	nivcsw	Numeric
26	min_flt	Numeric
27	maj_flt	Numeric

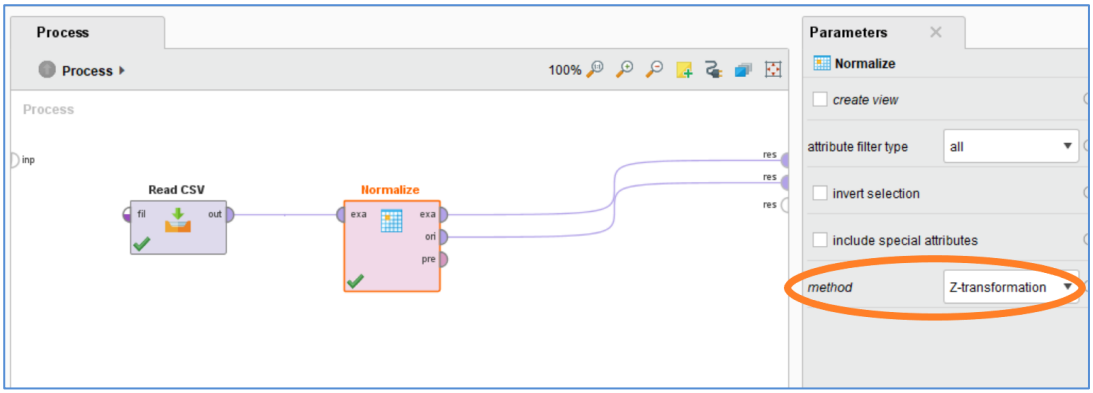
No	Attribut	Type Data
28	fs_excl_counter	Numeric
29	lock	Numeric
30	utime	Numeric
31	stime	Numeric
32	gtime	Numeric
33	cgtime	Numeric
34	signal_nvcsw	Numeric

Tabel 3.2 Contoh Isi Dataset Malware

millisecond	classification	state	usage_counter	prio	static_prio	normal_prio	policy	vm_pgoff	vm_truncate_count
995	malware	12288	0	3069378560	14274	0	0	0	13710
996	malware	12288	0	3069378560	14274	0	0	0	13710
997	malware	12288	0	3069378560	14274	0	0	0	13710
998	malware	12288	0	3069378560	14274	0	0	0	13710
999	malware	12288	0	3069378560	14274	0	0	0	13710
0	benign	0	0	3069403136	16447	0	0	0	14739
1	benign	0	0	3069403136	16447	0	0	0	14739
2	benign	0	0	3069403136	16447	0	0	0	14739
3	benign	0	0	3069403136	16447	0	0	0	14739
4	benign	0	0	3069403136	16447	0	0	0	14739

3.4 Normalisasi Data

Tahap normalisasi merupakan tahapan pra-pemrosesan yang dilakukan sebagai penyesuaian terhadap kemunculan nilai kontinu dalam fitur dataset yang akan mempengaruhi hasil proses klasifikasi dengan Naïve Bayes dan Decision Tree. Proses dilakukan dengan membaca dataset malware yang telah diinputkan sebelumnya ke dalam tool RapidMiner, kemudian mulai dilakukan tahap normalisasi. Proses normalisasi disini dimaksudkan untuk merubah jenis skala pengukuran yang dari data numerik. Proses normalisasi dilakukan dengan teknik Z-Score menggunakan fungsi Z-transpormation yang ada pada tool Rapidminer seperti pada gambar dan tabel dibawah ini :



Gambar 3.3 Proses Normalisasi Data

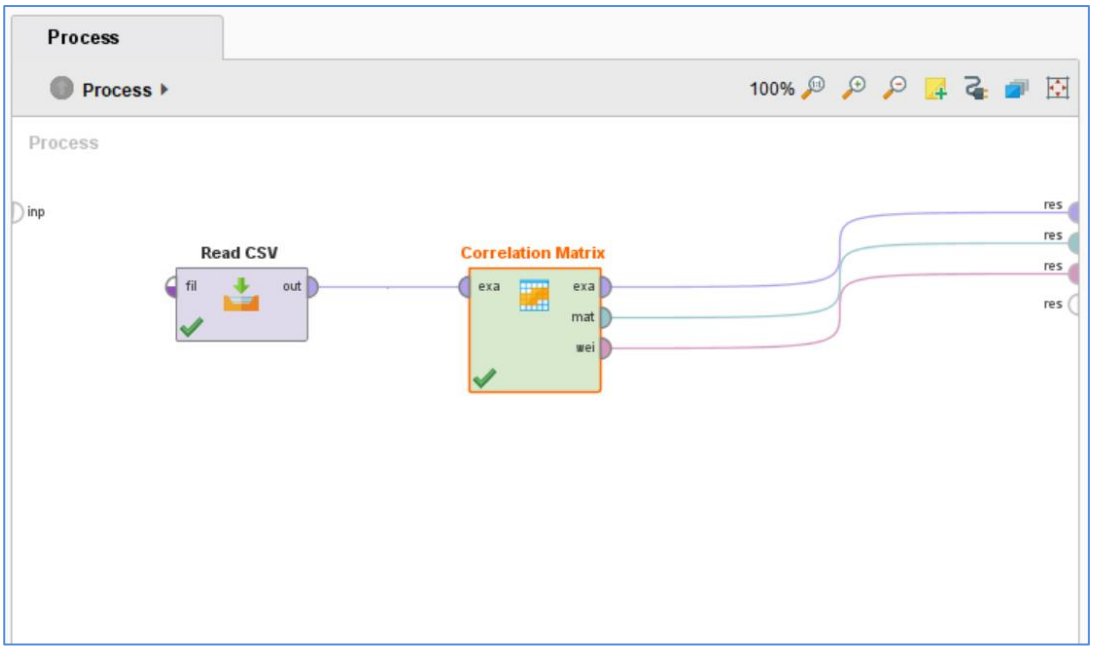
Tabel 3.3 Hasil Normalisasi Data

millisecond	classification	state	prio	static_prio	normal_prio	policy	vm_pgoff
-1.730311	malware	-0.16852	-1.10514	-0.848172634	0	0	0
-1.7025983	malware	-0.16852	-1.10514	-0.848172634	0	0	0
-1.6991342	malware	-0.16852	-1.10514	-0.848172634	0	0	0
-1.6159961	malware	-0.16852	-1.10514	-0.848172634	0	0	0
-1.612532	malware	-0.16852	-1.10514	-0.848172634	0	0	0
0.48324	benign	-0.16852	-1.39545	-0.506508685	0	0	0
0.49709634	benign	-0.16852	-1.39545	-0.506508685	0	0	0
0.50402452	benign	-0.16852	-1.39545	-0.506508685	0	0	0
0.55944989	benign	-0.16852	-1.39545	-0.506508685	0	0	0
0.56291398	benign	-0.16852	-1.39545	-0.506508685	0	0	0

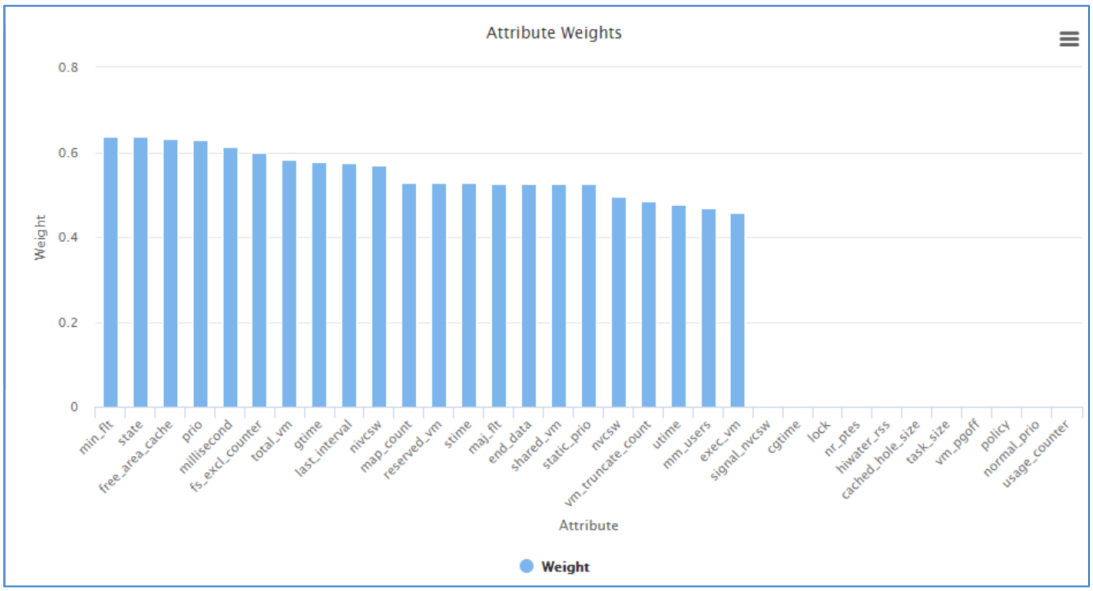
3.5 Korelasi Matrik

Tahapan pra-pemrosesan berikutnya adalah korelasi matrik yang digunakan untuk mereduksi dimensi dari suatu dataset dalam hal ini yang kita mereduksi atribut-atribut yang tidak kontributif. Tahapan ini dimaksudkan untuk mengurangi

dimensi dataset yang semula terdiri dari 34 atribut, setelah di lakukan operasi korelasi matrik dengan tool rapidminer hasilnya terdapat 11 atribut yang tidak kontributif sehingga hanya tersisa 23 atribut. Proses korelasi matrik seperti pada gambar dan tabel dibawah ini :



Gambar 3.4 Proses Korelasi Matrik



Gambar 3.5 Hasil Proses Korelasi Matrik

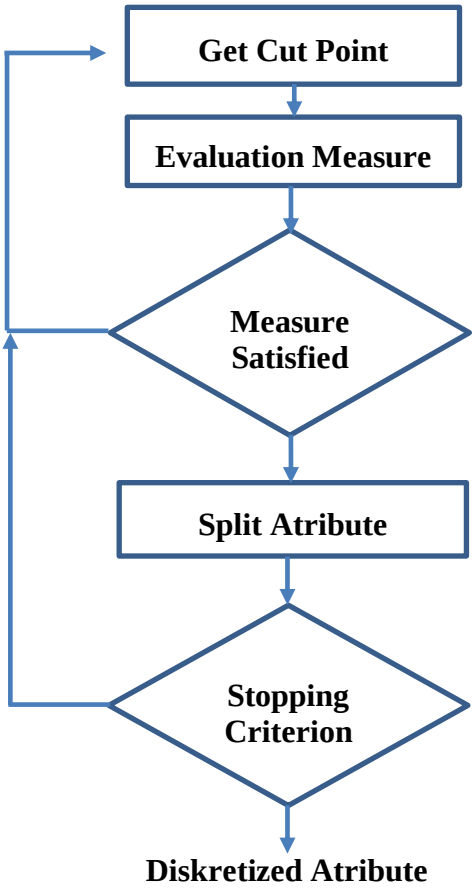
Setelah proses korelasi matrik dijalankan maka akan terlihat atribut-atribut yang menghasilkan nilai koefisien korelasi sama dengan 0 seperti terlihat pada gambar 3.4 di atas, sehingga atribut-atribut tersebut dapat kita hilangkan yaitu atribut signal_ncsw, cgtime, lock, nr_ptes, hiwater_rss, cached_hole_size, task_size, vm_pgoff, policy, normal_prio dan usage counter. Sehingga atribut yang tersisa yakni yang mempunyai nilai koefisien korelasi >0, seperti tabel dibawah ini :

Tabel 3.4 Dataset Hasil Korelasi Matrik

No	Attribut	Type Data
1.	millisecond	Numeric
2.	classification	String
3.	State	Numeric
4.	Prio	Numeric
5.	static_prio	Numeric
6.	vm_truncate_count	Numeric
7.	free_area_cache	Numeric
8.	mm_users	Numeric
9.	map_count	Numeric
10.	total_vm	Numeric
11.	shared_vm	Numeric
12.	exec_vm	Numeric
13.	reserved_vm	Numeric
14.	end_data	Numeric
15.	last_interval	Numeric
16.	nvcsw	Numeric
17.	nivcsw	Numeric
18.	min_flt	Numeric
19.	maj_flt	Numeric
20.	fs_excl_counter	Numeric
21.	Utime	Numeric
22.	Stime	Numeric
23.	Gtime	Numeric

3.6 Diskritisasi Data

Tahap pra-pemrosesan berikutnya adalah diskritisasi data. Teknik diskretisasi dapat digunakan untuk mereduksi sekumpulan nilai yang terdapat pada atribut continuous, dengan membagi range dari atribut ke dalam interval.



Gambar 3.6 Proses Diskretisasi secara umum

Proses diskretisasi secara umum terdiri dari 6 tahapan (gambar 3.3), yaitu:

a. Sorting, melakukan sorting nilai atribut continuous yang mau didiskretisasi.
b. Memilih kandidat “cut-point”, banyak fungsi evaluasi yang dapat digunakan seperti binning dan pengukuran entropy.
c. Melakukan perhitungan terhadap atribut-atribut yang akan didiskritisasi.
d. Mengecek hasil perhitungan jika no memilih kandidat cut point berikutnya sebaliknya jika ya dilanjutnya ke proses berikutnya.
e. Splitting, dilakukan evaluasi cut-point yang ada dan pilih satu yang terbaik dan lakukan split range nilai atribut continuous ke dalam dua partisi. Diskretisasi berlanjut untuk tiap partisi sampai kondisi berhenti tercapai.
f. Stopping criterion, diperlukan untuk menghentikan proses diskretisasi.

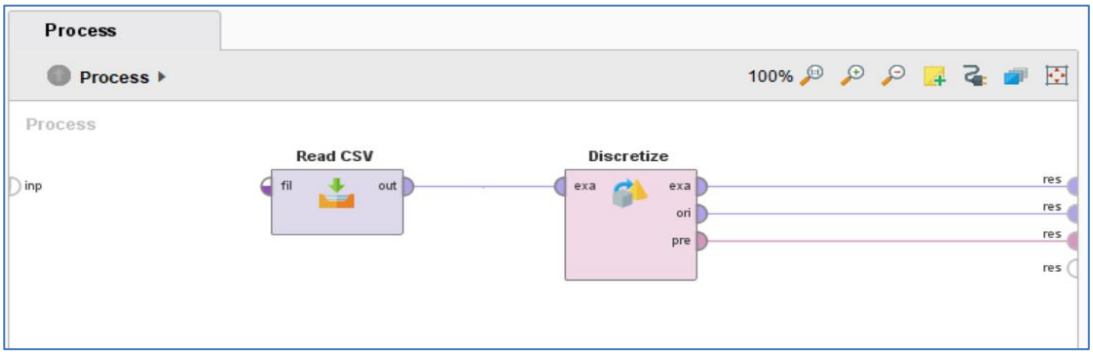
Terdapat 5 teknik untuk melakukan diskretisasi pada atribut continuous, yaitu: binning, cluster analysis, histogram analysis, entropy-based discretization, dan segmentation by “natural partitioning”.

Pada penelitian ini akan dicoba dengan menggunakan teknik binning yang akan diimplementasikan pada Tool Rapidminer, dimana teknik binning dilakukan dengan cara memeriksa “nilai tetangga”, yaitu nilai-nilai yang ada disekelilingnya.

Berikut adalah langkah-langkah teknik binning:

1) Data diurutkan dari yang terkecil sampai dengan yang terbesar.
2) Data yang sudah urut kemudian dipartisi ke dalam beberapa bin. Teknik partisi ke dalam bin ada 2 (dua) cara: equal-width (distance) partitioning dan equaldepth (frequency) partitioning.
3) Dilakukan smoothing dengan tiga macam teknik, yaitu: smoothing by binmeans, smoothing by bin-medians, dan smoothing by bin-boundaries.

Selanjutnya proses diskritisasi dari data hasil normalisasi pada tool RapidMiner bisa dilihat pada Gambar dibawah ini :



Gambar 3.7 Pemodelan Diskritisasi Data

Setelah proses diskritisasi dijalankan maka akan menghasilkan data yang telah diproses ke dalam pendiskritan. Adapun variabel yang didiskritisasi yaitu variabel millisecond, state, prio, static prio, vm_truncate_count, free_area_cache, mm_user, map_count, total_vm, shared_vm, exec_vm, reserved_vm, end_data, last interval, nvcsw, nivcsw, min_flt, maj_flt, fs_excl_counter, utime, stime, gtime. Contoh hasil diskritisasi bisa dilihat pada gambar di bawah ini :

classification	millisecond	state	prio	static_prio	vm_truncat...	free_area_c...	mm_users	map_count	total_vm	shared_vm
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]
malware	[-∞ - -0.6]	[-∞ - 15.2]	[-∞ - -0.6]	[-∞ - 0.4]	[-∞ - 0.1]	[-∞ - 3.5]	[-∞ - 0.1]	[-∞ - 0.6]	[-∞ - 2.2]	[-0.2 - ∞]

Gambar 3.8 Hasil Proses Diskritisasi Data

3.7 Modeling

Tahap ini juga dapat disebut tahap learning karena pada tahap ini data training diklasifikasikan oleh model dan kemudian menghasilkan sejumlah aturan. Pada penelitian ini, pembuatan model menggunakan dua algoritma, yaitu algoritma decision tree dan naïve bayes.

3.8 Evaluation

Pada tahap ini dilakukan pengujian terhadap model-model yang dikomparasi untuk mendapatkan informasi model yang paling akurat. Evaluasi dan validasi menggunakan metode cross validation, confusion matrix, dan kurva ROC.

3.9 Deployment

Setelah pembentukan model dan dilakukan analisa dan pengukuran pada tahap sebelumnya, selanjutnya pada tahap ini diterapkan model yang paling akurat untuk pendeteksian malware.