

ABSTRAK
IMPLEMENTASI METODE *RANDOM FOREST* DAN SELEKSI FITUR
***INFORMATION GAIN* PADA KLASIFIKASI POTENSI PENYAKIT**
JANTUNG

Qurrotul Aini

Program Studi Teknik Informatika, Fakultas Teknik

Universitas Trunojoyo Madura

Suatu kondisi tidak menular yang dapat menyebabkan penyakit jantung. Penyakit jantung merupakan salah satu penyakit tidak menular yang dapat menyerang tubuh secara tiba-tiba karena adanya penyempitan pada aliran darah menuju jantung. Pada penelitian ini menggunakan metode *random forest* untuk mengklasifikasi penyakit jantung, dimana jumlah datanya yaitu 1190, 11 fitur, kelas target ada 2 (penyakit jantung dan normal). Penelitian ini bertujuan menghitung dan mengetahui hasil ketepatan tingkat akurasi antara *dataset* dengan teknik seleksi fitur dan *dataset* tanpa seleksi fitur dari pemodelan yang dibuat menggunakan metode *random forest*. Hasil akurasi menggunakan jumlah fitur sebanyak 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (teknik seleksi fitur) yaitu 80.26%, 81.12%, 83.26%, 83.69%, 83.26%, 86.70%, 88.84%, 88.84%, 85.84%, 90.56% pada percobaan tanpa tuning parameter dengan *split validation*, sedangkan tanpa teknik seleksi fitur mencapai 91.42% (akurasi optimum). Hasil akurasi menggunakan jumlah fitur sebanyak 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (teknik seleksi fitur) yaitu 78.66%, 76.94%, 81.33%, 81.33%, 81.84%, 83.65%, 86.66%, 89.16%, 88.64%, 91.05% pada percobaan tanpa tuning parameter dengan *5-fold validation*, sedangkan tanpa teknik seleksi fitur mencapai 91.22% (akurasi optimum). Hasil akurasi dengan jumlah fitur sebanyak 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (teknik seleksi fitur) yaitu 80.26%, 81.12%, 83.26%, 83.69%, 83.97%, 85.84%, 87.98%, 88.84%, 87.55%, 90.56% pada percobaan menggunakan hasil tuning parameter dengan *split validation*, sedangkan tanpa teknik seleksi fitur mencapai 91.42% (akurasi optimum). Hasil akurasi dengan jumlah fitur 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (teknik seleksi fitur) yaitu 78.66%, 77.45%, 81.41%, 81.67%, 81.58%, 84.25%, 86.23%, 89.07%, 88.47% dan 90.71% pada percobaan

menggunakan hasil tuning parameter dengan *5-fold validation*, sedangkan tanpa teknik seleksi fitur mencapai 90.88% (akurasi optimum). Dari hasil skenario uji coba, akurasi optimum mencapai 91.42% pada pemodelan tanpa seleksi fitur menggunakan hasil tuning parameter dan tanpa tuning parameter dengan *split validation*. Hasil penelitian ini menunjukkan bahwa teknik seleksi fitur mencapai akurasi yang lebih rendah daripada menggunakan pemodelan tanpa seleksi fitur. Hal ini bisa disebabkan karena adanya pertimbangan fitur dominan dan tidak dominan, serta teknik seleksi fitur dapat berpengaruh jika diterapkan pada data dengan jumlah *record* dan fitur yang banyak.

Kata kunci : penyakit jantung, klasifikasi, *random forest*

ABSTRACT
IMPLEMENTATION OF RANDOM FOREST METHOD AND FEATURES
SELECTION INFORMATION GAIN TO CLASSIFICATION OF
POTENTIAL HEART DISEASE

Qurrotul Aini

Department of Informatics Engineering, Faculty of Engineering
Trunojoyo University of Madura

Heart disease is a non-communicable disease that can attack the body suddenly due to narrowing of the blood flow to the heart. In this study using the random forest method to classify heart disease, where the amount of data is 1190, 11 features, there are 2 target classes (heart disease and normal). This study aims to calculate and find out the results of the accuracy level between datasets with feature selection techniques and datasets without feature selection from modeling made using the random forest method. Accuracy results using the number of features as much as 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (feature selection technique) namely 80.26%, 81.12%, 83.26%, 83.69%, 83.26%, 86.70%, 88.84 %, 88.84%, 85.84%, 90.56% in the experiment without parameter tuning with split validation, while without the feature selection technique it reached 91.42% (optimum accuracy). Accuracy results using the number of features of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (feature selection technique) namely 78.66%, 76.94%, 81.33%, 81.33%, 81.84%, 83.65%, 86.66 %, 89.16%, 88.64%, 91.05% in the experiment without parameter tuning with 5-fold validation, while without feature selection technique it reached 91.22% (optimum accuracy). Accuracy results with a number of features of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (feature selection technique) namely 80.26%, 81.12%, 83.26%, 83.69%, 83.97%, 85.84%, 87.98 %, 88.84%, 87.55%, 90.56% in the experiment using the results of tuning parameters with split validation, while without feature selection techniques it reached 91.42% (optimum accuracy). Accuracy results with the number of features 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (feature selection technique) namely 78.66%, 77.45%, 81.41%, 81.67%, 81.58%, 84.25%, 86.23% , 89.07%, 88.47%

and 90.71% in the experiment using parameter tuning results with 5-fold validation, while without feature selection techniques it reached 90.88% (optimum accuracy). From the results of the trial scenario, the optimum accuracy reached 91.42% in modeling without feature selection using the results of parameter tuning and without parameter tuning with split validation. The results of this study indicate that the feature selection technique achieves lower accuracy than using modeling without feature selection. This can be due to the consideration of dominant and non-dominant features, and feature selection techniques can have an effect if applied to data with a large number of records and features.

Keywords : *heart disease, classification, random forest*

DAFTAR ISI

COVER

HALAMAN JUDUL	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN ORISINALITAS	iv
KATA PENGANTAR.....	v
DAFTAR ISI.....	vii
DAFTAR GAMBAR.....	x
DAFTAR TABEL	xiv
ABSTRAK	xv
ABSTRACT	xvii
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah.....	3
1.2.1. Permasalahan	3
1.2.2. Metode Usulan.....	4
1.2.3. Pertanyaan Penelitian.....	4
1.3. Batasan Masalah.....	4
1.4. Tujuan Penelitian	4
1.5. Manfaat Penelitian	5
1.6. Sistematika Penulisan	5
BAB II KAJIAN PUSTAKA	7
2.1. Penelitian Terdahulu.....	7
2.2. Penyakit Jantung.....	13
2.3. <i>Data Mining</i>	14
2.4. <i>Machine Learning</i>	14
2.5. <i>Missing Value</i>	15
2.6. <i>Outlier</i>	16
2.7. Diskritisasi.....	16
2.8. Seleksi Fitur <i>Information Gain</i>	17
2.9. Klasifikasi.....	18
2.10. <i>Random Forest</i>	19

2.11.	<i>Python</i>	21
2.12.	<i>Tuning Parameter Menggunakan Random Search CV</i>	22
2.13.	<i>Evaluasi Model</i>	22
BAB III METODE USULAN		26
3.1.	Kebutuhan Sistem	26
3.2.	Tahapan Penelitian	26
3.3.	Data Collection	29
3.4.	Pengelolaan Data Awal	30
3.5.	Tahapan Seleksi Fitur	31
3.6.	Tahapan Metode Klasifikasi <i>Random Forest</i>	32
3.7.	Evaluasi Model	34
3.8.	Skenario Uji coba	34
BAB IV HASIL DAN PEMBAHASAN		36
4.1.	Hasil Proses Pengelolaan Data Awal	36
4.1.1.	Hasil Proses <i>Import Library</i>	36
4.1.2.	Hasil Proses <i>Load Dataset</i>	37
4.1.3.	Hasil Proses Imputasi <i>Missing Value</i>	40
4.1.4.	Hasil Proses Eksplorasi Data	43
4.1.5.	Hasil Proses Deteksi <i>Outlier</i>	50
4.1.6.	Hasil Proses Diskritisasi	52
4.2.	Hasil Proses Seleksi Fitur <i>Information Gain</i>	54
4.3.	Hasil Proses <i>Split Data</i>	56
4.4.	Hasil Proses Tuning Parameter <i>Random SearchCV</i>	57
4.5.	Hasil Pemodelan <i>Random Forest</i> dan Proses Skenario Ujicoba	59
4.5.1.	Pemodelan <i>Random Forest</i> Tanpa Teknik Seleksi Fitur	59
4.5.1.1.	Pemodelan <i>Random Forest</i> Tanpa Tuning Parameter Menggunakan <i>Split Validation</i>	59
4.5.1.2.	Pemodelan <i>Random Forest</i> Tanpa Tuning Parameter Menggunakan <i>5-Fold Validation</i>	62
4.5.1.3.	Pemodelan <i>Random Forest</i> Menggunakan Hasil Tuning Parameter dengan <i>Split Validation</i>	63

4.5.1.4.	Pemodelan <i>Random Forest</i> Menggunakan Hasil Tuning Parameter 5-Fold Validation	64
4.5.2.	Pemodelan <i>Random Forest</i> Menggunakan Teknik Seleksi Fitur	65
4.5.2.1.	Pemodelan Menggunakan 10 Fitur	65
4.5.2.2.	Pemodelan Menggunakan 9 Fitur	69
4.5.2.3.	Pemodelan Menggunakan 8 Fitur	72
4.5.2.4.	Pemodelan Menggunakan 7 Fitur	76
4.5.2.5.	Pemodelan Menggunakan 6 Fitur	79
4.5.2.6.	Pemodelan Menggunakan 5 Fitur	83
4.5.2.7.	Pemodelan Menggunakan 4 Fitur	86
4.5.2.8.	Pemodelan Menggunakan 3 Fitur	90
4.5.2.9.	Pemodelan Menggunakan 2 Fitur	93
4.5.2.10.	Pemodelan Menggunakan 1 Fitur.....	97
4.5.3.	Hasil Analisis Skenario Uji coba.....	100
BAB V	PENUTUP	104
5.1.	Kesimpulan	104
5.2.	Saran.....	105
REFERENSI		106
LAMPIRAN		110
BIOGRAFI PENULIS		145

DAFTAR GAMBAR

Gambar 2.1 Struktur <i>random forest</i>	20
Gambar 2.2 Ilustrasi <i>5-fold cross validation</i>	23
Gambar 3.1 Tahapan penelitian.....	27
Gambar 3.2 <i>Flowchart</i> metode <i>random forest</i>	33
Gambar 4. 1 <i>Library</i>	36
Gambar 4. 2 <i>Code load dataset</i>	37
Gambar 4. 3 <i>Toy dataset</i>	38
Gambar 4. 4 <i>Code rename</i> fitur	38
Gambar 4. 5 Distribusi fitur <i>chest pain type</i>	39
Gambar 4. 6 Distribusi fitur <i>rest ecg</i>	39
Gambar 4. 7 Distribusi fitur <i>st_slope</i>	39
Gambar 4. 8 Distribusi fitur <i>sex</i>	40
Gambar 4. 9 <i>Missing value</i> pada fitur <i>st slope</i>	40
Gambar 4. 10 <i>Data missing value</i> setelah di <i>replace Nan</i>	41
Gambar 4. 11 Data pada fitur target kelas 1	41
Gambar 4. 12 <i>Code</i> imputasi modus	42
Gambar 4. 13 Hasil proses imputasi modus	42
Gambar 4. 14 Distribusi jumlah fitur <i>st slope</i>	43
Gambar 4. 15 <i>Plot missing value</i>	43
Gambar 4. 16 Distribusi data penyakit jantung pada kelas target	44
Gambar 4. 17 Distribusi data pada fitur <i>sex</i> dan <i>age</i>	44
Gambar 4. 18 Distribusi data pada fitur <i>sex</i> dan <i>age</i> kelas 0	45
Gambar 4. 19 Distribusi data pada fitur <i>sex age</i> kelas 1	45
Gambar 4. 20 Distribusi data pada fitur <i>chest pain type</i>	46
Gambar 4. 21 Jumlah distribusi data pada fitur <i>chest pain type</i>	46
Gambar 4. 22 Distribusi data pada fitur <i>resting ecg</i>	47
Gambar 4. 23 Jumlah distribusi data pada fitur <i>resting ecg</i>	47
Gambar 4. 24 Distribusi data pada fitur <i>st slope</i>	48
Gambar 4. 25 Jumlah distribusi data pada fitur <i>st slope</i>	48
Gambar 4. 26 Distribusi data pada fitur numerik	49

Gambar 4. 27	Korelasi data	50
Gambar 4. 28	Data pada fitur numerik	51
Gambar 4. 29	Hasil perhitungan <i>z-score</i>	51
Gambar 4. 30	Hasil data <i>outlier</i>	52
Gambar 4. 31	Jumlah distribusi data pada kelas target	52
Gambar 4. 32	Data sebelum proses diskritisasi	53
Gambar 4. 33	Hasil diskritisasi	54
Gambar 4. 34	Nilai <i>evaluasi</i> kinerja pemodelan <i>Random Forest</i>	61
Gambar 4. 35	<i>Confusion matrix</i> pemodelan1	62
Gambar 4. 36	Hasil evaluasi kinerja pemodelan <i>kfold1</i>	63
Gambar 4. 37	Nilai <i>evaluasi</i> kinerja pemodelan tuning parameter	63
Gambar 4. 38	<i>Confusion matrix</i> pemodelan2	64
Gambar 4. 39	Hasil evaluasi kinerja pemodelan <i>kfold2</i>	65
Gambar 4. 40	Nilai evaluasi pemodelan 10 fitur tanpa tuning	66
Gambar 4. 41	<i>Confusion matrix</i> 10 fitur tanpa tuning parameter.....	66
Gambar 4. 42	Hasil evaluasi pemodelan 10 fitur <i>kfold</i> tanpa tuning	67
Gambar 4. 43	Nilai evaluasi pemodelan 10 fitur tuning	67
Gambar 4. 44	<i>Confusion matrix</i> 10 fitur tuning parameter	68
Gambar 4. 45	Hasil evaluasi pemodelan 10 fitur <i>kfold</i> tuning.....	68
Gambar 4. 46	Nilai evaluasi pemodelan 9 fitur tanpa tuning	69
Gambar 4. 47	<i>Confusion matrix</i> 9 fitur tanpa tuning parameter.....	70
Gambar 4. 48	Hasil evaluasi pemodelan 9 fitur <i>kfold</i> tanpa tuning	70
Gambar 4. 49	Nilai evaluasi pemodelan 9 fitur tuning	71
Gambar 4. 50	<i>Confusion matrix</i> 9 fitur tuning parameter	71
Gambar 4. 51	Hasil evaluasi pemodelan 9 fitur <i>kfold</i> tuning.....	72
Gambar 4. 52	Nilai evaluasi pemodelan 8 fitur tanpa tuning	72
Gambar 4. 53	<i>Confusion matrix</i> 8 fitur tanpa tuning parameter.....	73
Gambar 4. 54	Hasil evaluasi pemodelan 8 fitur <i>kfold</i> tanpa tuning	74
Gambar 4. 55	Nilai evaluasi pemodelan 8 fitur tuning	74
Gambar 4. 56	<i>Confusion matrix</i> 8 fitur tuning parameter	75
Gambar 4. 57	Hasil evaluasi pemodelan 8 fitur <i>kfold</i> tuning.....	75
Gambar 4. 58	Nilai evaluasi pemodelan 7 fitur tanpa tuning	76

Gambar 4. 59	<i>Confusion matrix</i> 7 fitur tanpa tuning parameter.....	77
Gambar 4. 60	Hasil evaluasi pemodelan 7 fitur kfold tanpa tuning	77
Gambar 4. 61	Nilai evaluasi pemodelan 7 fitur tuning	78
Gambar 4. 62	<i>Confusion matrix</i> 7 fitur tuning parameter	78
Gambar 4. 63	Hasil evaluasi pemodelan 7 fitur kfold tuning.....	79
Gambar 4. 64	Nilai evaluasi pemodelan 6 fitur tanpa tuning	79
Gambar 4. 65	<i>Confusion matrix</i> 6 fitur tanpa tuning parameter.....	80
Gambar 4. 66	Hasil evaluasi pemodelan 6 fitur kfold tanpa tuning	81
Gambar 4. 67	Nilai evaluasi pemodelan 6 fitur tuning	81
Gambar 4. 68	<i>Confusion matrix</i> 6 fitur tuning parameter	82
Gambar 4. 69	Hasil evaluasi pemodelan 6 fitur kfold tuning.....	82
Gambar 4. 70	Nilai evaluasi pemodelan 5 fitur tanpa tuning	83
Gambar 4. 71	<i>Confusion matrix</i> 5 fitur tanpa tuning parameter.....	84
Gambar 4. 72	Hasil evaluasi pemodelan 5 fitur kfold tanpa tuning	84
Gambar 4. 73	Nilai evaluasi pemodelan 5 fitur tuning	85
Gambar 4. 74	<i>Confusion matrix</i> 10 fitur tuning parameter	85
Gambar 4. 75	Hasil evaluasi pemodelan 5 fitur kfold tuning.....	86
Gambar 4. 76	Nilai evaluasi pemodelan 4 fitur tanpa tuning	86
Gambar 4. 77	<i>Confusion matrix</i> 4 fitur tanpa tuning parameter.....	87
Gambar 4. 78	Hasil evaluasi pemodelan 4 fitur kfold tanpa tuning	88
Gambar 4. 79	Nilai evaluasi pemodelan 4 fitur tuning	88
Gambar 4. 80	<i>Confusion matrix</i> 4 fitur tuning parameter	89
Gambar 4. 81	Hasil evaluasi pemodelan 4 fitur kfold tuning.....	89
Gambar 4. 82	Nilai evaluasi pemodelan 3 fitur tanpa tuning	90
Gambar 4. 83	<i>Confusion matrix</i> 3 fitur tanpa tuning parameter.....	91
Gambar 4. 84	Hasil evaluasi pemodelan 3 fitur kfold tanpa tuning	91
Gambar 4. 85	Nilai evaluasi pemodelan 3 fitur tuning	92
Gambar 4. 86	<i>Confusion matrix</i> 3 fitur tuning parameter	92
Gambar 4. 87	Hasil evaluasi pemodelan 3 fitur kfold tuning.....	93
Gambar 4. 88	Nilai evaluasi pemodelan 2 fitur tanpa tuning	93
Gambar 4. 89	<i>Confusion matrix</i> 2 fitur tanpa tuning parameter.....	94
Gambar 4. 90	Hasil evaluasi pemodelan 2 fitur kfold tanpa tuning	95

Gambar 4. 91	Nilai evaluasi pemodelan 2 fitur tuning	95
Gambar 4. 92	<i>Confusion matrix</i> 2 fitur tuning parameter	96
Gambar 4. 93	Hasil evaluasi pemodelan 2 fitur kfold tuning.....	96
Gambar 4. 94	Nilai evaluasi pemodelan 1 fitur tanpa tuning	97
Gambar 4. 95	<i>Confusion matrix</i> 1 fitur tanpa tuning parameter.....	98
Gambar 4. 96	Hasil evaluasi pemodelan 1 fitur kfold tanpa tuning	98
Gambar 4. 97	Nilai evaluasi pemodelan 1 fitur tuning	99
Gambar 4. 98	<i>Confusion matrix</i> 1 fitur tuning parameter	99
Gambar 4. 99	Hasil evaluasi pemodelan 1 fitur kfold tuning.....	100
Gambar 4. 100	Grafik akurasi tanpa tuning parameter <i>split validation</i>	102
Gambar 4. 101	Grafik akurasi tanpa tuning parameter <i>5-fold validation</i>	102
Gambar 4. 102	Grafik akurasi tuning parameter <i>split validation</i>	103
Gambar 4. 103	Grafik akurasi tuning parameter <i>5-fold validation</i>	103

DAFTAR TABEL

Tabel 2.1 Rangkuman Jurnal Penelitian Terdahulu	9
Tabel 2.2 Parameter metode <i>random forest</i>	22
Tabel 2.3 Tabel <i>confusion matrix</i>	23
Tabel 2.4 Tabel keakuratan AUC	25
Tabel 3.1 Penjelasan fitur <i>dataset</i> penyakit jantung.....	29
Tabel 3. 2 Skenario uji coba	34
Tabel 4. 1 <i>Code split</i> kolom fitur dan target.....	53
Tabel 4. 2 <i>Code</i> diskritisasi	53
Tabel 4. 3 <i>Code</i> seleksi fitur <i>information gain</i>	54
Tabel 4. 4 Hasil seleksi fitur <i>information gain</i>	55
Tabel 4. 5 <i>Code split data</i>	56
Tabel 4. 6 Rincian hasil <i>split data</i>	56
Tabel 4. 7 <i>Code</i> cek distribusi data <i>training</i> dan <i>testing</i> pada fitur target.....	56
Tabel 4. 8 Hasil distribusi jumlah data <i>training</i> dan <i>testing</i> fitur target.....	57
Tabel 4. 9 <i>Code</i> distribusi parameter.....	57
Tabel 4. 10 <i>Code</i> tuning parameter <i>random searchcv</i>	58
Tabel 4. 11 Hasil tuning parameter <i>random searchcv</i>	58
Tabel 4. 12 <i>Code</i> pemodelan <i>random forest</i>	60
Tabel 4. 13 <i>Code save</i> hasil evaluasi pemodelan	61
Tabel 4. 14 Hasil akurasi pemodelan tanpa seleksi fitur dan seleksi fitur	101