JNATIA Volume 1, Nomor 4, Agustus 2023 p-ISSN: 2986-3929
Jurnal Nasional Teknologi Informasi dan Aplikasinya

Implementasi Random Forest dengan LASSO Dalam Klasifikasi Penyakit yang Ditularkan Melalui Nyamuk

Kadek Dwitya Adhi Pradyto[a1], Made Agung Raharja[a2]

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali Indonesia
[1]adhipradyto32@gmail.com
[2]made.agung@unud.ac.id

Abstract

Several diseases that can attack human health can be transmitted through disease vectors. One of the insects belonging to the disease vector is the mosquito. Diseases that can attack humans due to transmission through mosquitoes include malaria, dengue fever, chikungunya, yellow fever, rift valley fever, and many more. With so many types of diseases that are transmitted by mosquitoes and the symptoms that look quite similar, a classification process is carried out to distinguish the types of diseases. In this study, the classification was carried out using the Random Forest algorithm with the LASSO algorithm for feature selection. It was found that the average accuracy values of the Random Forest before and after carrying out feature selection using LASSO were 88% and 76%, respectively. From the values obtained, it can be concluded that the Random Forest has better performance without feature selection using the LASSO method.

Keywords: *Classification, Random Forest, LASSO, Mosquito-Borne Diseases*

1. Pendahuluan

Beberapa penyakit yang bisa menyerang kesehatan manusia dapat ditularkan melalui vektor penyakit. Vektor penyakit merupakan jenis serangga yang bisa menularkan penyakit. Terjadi pertumbuhan sumber penyakit seperti virus di dalam tubuh vektor dengan jumlah yang cukup untuk menimbulkan penyakit baru [1]. Salah satu serangga yang tergolong ke dalam vektor penyakit adalah nyamuk. Nyamuk betina menghisap darah manusia untuk dijadikan sumber energi bagi perkembangan telurnya. Melalui proses menghisap darah inilah nyamuk dapat menularkan penyakit yang mereka bawa. Penyakit-penyakit yang dapat menyerang manusia akibat penularan melalui nyamuk antara lain yaitu malaria, demam berdarah *dengue* (DBD), *chikungunya*, demam kuning, demam *rift valley*, dan masih banyak lagi.

Dengan banyaknya jenis penyakit yang ditularkan melalui nyamuk serta gejala yang terlihat cukup mirip yaitu adanya demam, maka untuk membedakan jenis-jenis penyakit tersebut dilakukan proses klasifikasi. Klasifikasi merupakan proses pengelompokkan sekumpulan kelas berdasarkan data-data yang ada [2]. Klasifikasi dapat digunakan untuk mengelompokkan dua jenis kelas (*binary class*) ataupun lebih dari dua jenis kelas (*multi-class*). Salah satu algoritma yang bisa digunakan baik untuk klasifikasi *binary class* ataupun *multi-class* yaitu *Random Forest*. Algoritma *Random Forest* merupakan pengembangan dari *Classification and Regression Tree* (CART) dengan menerapkan *random feature selection* dan *bagging* (*bootstrap-aggregating*). Model klasifikasi dari algoritma *Random Forest* merupakan sekumpulan *Decision Tree*. Kelemahan dari penggunaan *Decision Tree* dalam proses klasifikasi yaitu terbentuknya model yang *overfitting* karena terpengaruh *noise* dari data latih. *Overfitting* adalah suatu kondisi dimana model yang dilatih lebih condong dalam memprediksi data latih daripada data uji. Dalam mengatasi *overfitting*, diterapkan algoritma *Random Forest* yang terdiri dari sekumpulan *Decision Tree* yang dilatih dengan keadaan yang berbeda di tiap pohonnya dan masih bisa mendapatkan akurasi yang maksimum [3].

Terdapat banyak cara yang bisa dilakukan dalam mengoptimalkan proses klasifikasi salah satunya yaitu dengan menyeleksi fitur yang tidak ada hubungannya dengan kelas prediksi. Proses seleksi fitur akan menghapus variabel yang tidak bisa diprediksi dan berlebihan. Proses klasifikasi dengan fitur yang optimal akan menghasilkan model klasifikasi yang bekerja lebih cepat serta dapat mengurangi *overfitting* [4].

Pada penelitian ini, dilakukan proses klasifikasi penyakit yang ditularkan melalui nyamuk menggunakan algoritma *Random Forest*. Adapun optimasi yang dilakukan yaitu penyeleksian fitur dengan menggunakan LASSO. Penelitian ini juga akan menguji tingkat akurasi *Random Forest* sebelum dan sesudah seleksi fitur menggunakan LASSO.

2. Metode Penelitian

2.1. Teknik Pengumpulan Data

Data yang akan digunakan dalam penelitian ini merupakan data sekunder yang didapat dari *website* Kaggle (https://www.kaggle.com/datasets/richardbernat/vector-borne-disease-prediction). Data pada Kaggle masih berupa data penyakit yang ditularkan melalui vektor penyakit. Maka dari itu, data akan dibatasi sampai baris ke-184 yang merupakan bagian data penyakit yang ditularkan melalui nyamuk.

	sudden_fever	headache	mouth_bleed	nose_bleed	muscle_pain	...	ulcers	toenail_loss	speech_problem	bullseye_rash	prognosis
0	0	1	1	1	1	...	0	0	0	0	Chikungunya
1	1	1	1	1	1	...	0	0	0	0	Chikungunya
2	0	1	0	1	0	...	0	0	0	0	Chikungunya
3	0	0	0	0	0	...	0	0	0	0	Chikungunya
4	1	0	0	0	1	...	0	0	0	0	Chikungunya
...	...	...	...	...	...	...	...	...	...	...	...
179	1	0	0	1	1	...	0	0	0	0	West Nile fever
180	0	1	0	0	0	...	0	0	0	0	West Nile fever
181	0	0	1	1	1	...	0	0	0	0	West Nile fever
182	1	0	0	1	0	...	0	0	0	0	West Nile fever
183	0	0	1	1	1	...	0	0	0	0	West Nile fever

184 rows × 65 columns

Gambar 1. Data Penyakit Yang Ditularkan Oleh Nyamuk

Pada data tersebut, terdapat 64 jenis fitur untuk mendeteksi penyakit yang ditularkan melalui nyamuk. Fitur-fitur tersebut terdiri dari "*sudden_fever*" sampai "*bullseye_rash*". Terdapat 8 kelas yang akan diklasifikasikan. Kelas-kelas tersebut antara lain "*Chikungunya*", "*Dengue*", "*Rift Valley fever*", "*Yellow Fever*", "*Zika*", "*Malaria*", "*Japanese encephalitis*", dan "*West Nile fever*" Total keseluruhan data yang digunakan sebanyak 184 data dengan pembagian data latih dan data uji adalah 70:30.

2.2. Seleksi Fitur Menggunakan LASSO

LASSO adalah singkatan dari *Least Absolute Shrinkage and Selection Operator* dan merupakan metode yang cukup bagus dalam melakukan seleksi fitur. Metode ini bekerja dengan proses penyusutan nilai koefisien dari fitur yang sedang diuji menjadi nol. Setelah itu, metode ini memilih variabel-variabel yang tidak bernilai nol untuk menjadi fitur yang digunakan untuk proses klasifikasi [4].

2.3. Random Forest

Random Forest merupakan salah satu algoritma *Machine Learning* yang bekerja dengan mengombinasikan sejumlah algoritma *Decision Tree* dalam pengambilan keputusannya [5].

Metode ini digunakan untuk membangun *Decision Tree* yang terdiri dari *root node*, *internal node*, dan *leaf node* dengan mengambil fitur secara acak (*random*). *Root node* merupakan bagian yang terletak paling atas, atau biasa disebut sebagai akar dari pohon. *Internal node* adalah bagian percabangan, dimana bagian ini memiliki keluaran minimal dua dan hanya ada satu masukkan. Sedangkan *leaf node* merupakan bagian terakhir yang hanya memiliki satu masukkan dan tidak memiliki keluaran [6].

Proses klasifikasi menggunakan algoritma *Random Forest* akan dilakukan sebelum dan sesudah melakukan seleksi fitur dengan LASSO. Pencarian model terbaik dari *Random Forest* akan dilakukan dengan menggunakan metode *Random Search*. Metode ini digunakan karena lebih unggul dibandingkan dengan *Grid Search* dan *Bayesian Search* dalam mencari *parameter* dari *Random Forest* [5]. Adapun *parameter* yang dicari menggunakan *Random Search* yaitu "*n_estimators*", "*criterion*", "*max_depth*", "*min_samples_split*", dan "*max_features*".

2.4. Evaluasi

Evaluasi dilakukan dengan menghitung nilai akurasi dan presisi dari model. Pengujian akurasi dilakukan untuk mengetahui seberapa akurat model dalam melakukan prediksi. Sedangkan pengujian presisi dilakukan untuk mengetahui seberapa benar model melakukan prediksi di tiap kelasnya. Nilai akurasi dan presisi dapat ditentukan dengan rumus berikut.

$$Akurasi = \frac{Prediksi\ benar}{\text{Total prediksi}} \quad (1)$$

$$Presisi = \frac{True\ positive}{True\ positive + False\ positive} \quad (2)$$

3. Hasil dan Pembahasan.

3.1. Seleksi Fitur

Proses seleksi fitur, nilai *alpha* yang digunakan pada LASSO adalah 0,1. Adapun nilai koefisien setiap fitur yang didapat setelah melakukan seleksi fitur dengan metode LASSO dapat dilihat pada Gambar 2 di bawah.

```
array([0.14982622, 0.        , 0.        , 0.34365757, 0.26094839,
       0.        , 0.04326063, 0.        , 0.        , 0.        ,
       0.        , 0.06598752, 0.        , 0.        , 0.        ,
       0.        , 0.03846259, 0.        , 0.        , 0.        ,
       0.16111974, 0.0686008 , 0.33595216, 0.05387347, 0.        ,
       0.046126  , 0.3407998 , 0.23243593, 0.        , 0.        ,
       0.15462936, 0.        , 0.20797139, 0.99854777, 0.85869685,
       0.32251722, 0.14780409, 0.21390347, 0.4914114 , 0.        ,
       0.13825577, 0.62696595, 0.29986494, 0.        , 0.        ,
       0.09864327, 0.35127592, 0.        , 0.        , 0.        ,
       0.        , 0.05747193, 0.        , 0.        , 0.58092856,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        ])
```

Gambar 2. Nilai Koefisien Setiap Fitur

Fitur-fitur yang akan digunakan dalam proses klasifikasi memiliki nilai koefisien yang lebih dari nol (koefisien > 0). Berdasarkan nilai koefisien yang didapat di atas, daftar fitur-fitur yang terpilih dan tidak terpilih dapat dilihat pada Gambar 3 dan Gambar 4.

```
array(['sudden_fever', 'nose_bleed', 'muscle_pain', 'vomiting', 'ascites',
       'myalgia', 'stomach_pain', 'orbital_pain', 'neck_pain', 'weakness',
       'weight_loss', 'gum_bleed', 'jaundice', 'inflammation',
       'loss_of_appetite', 'urination_loss', 'slow_heart_rate',
       'abdominal_pain', 'light_sensitivity', 'yellow_skin',
       'yellow_eyes', 'microcephaly', 'rigor', 'bitter_tongue',
       'cocacola_urine', 'hypoglycemia', 'confusion', 'lymph_swells'],
      dtype='<U21')
```

Gambar 3. Hasil Seleksi Fitur Yang Terpilih

```
array(['headache', 'mouth_bleed', 'joint_pain', 'rash', 'diarrhea',
       'hypotension', 'pleural_effusion', 'gastro_bleeding', 'swelling',
       'nausea', 'chills', 'digestion_trouble', 'fatigue', 'skin_lesions',
       'back_pain', 'coma', 'diziness', 'red_eyes', 'facial_distortion',
       'convulsion', 'anemia', 'prostraction', 'hyperpyrexia',
       'stiff_neck', 'irritability', 'tremor', 'paralysis',
       'breathing_restriction', 'toe_inflammation', 'finger_inflammation',
       'lips_irritation', 'itchiness', 'ulcers', 'toenail_loss',
       'speech_problem', 'bullseye_rash'], dtype='<U21')
```

Gambar 4. Hasil Seleksi Fitur Yang Tidak Terpilih

3.2. Hasil Klasifikasi

Pada penelitian ini, proses klasifikasi dilakukan sebanyak 2 kali yaitu sebelum melakukan seleksi fitur dan sesudah melakukan seleksi fitur. Sebelum melakukan klasifikasi dengan *Random Forest*, dicari model terbaik dengan metode *Random Search*. Adapun nilai-nilai dari *hyperparameter* yang akan dicari bisa dilihat pada Tabel 1.

Tabel 1. Nilai *Hyperparameter*

Nama *Parameter*	Nilai
n_estimators	[50, 100, 150, 200]
criterion	[‘gini’, ‘entropy’]
max_depth	[None, 5, 10]
min_samples_split	[2, 4, 6, 8, 10]
max_features	['sqrt', 'log2']

Pada saat sebelum melakukan seleksi fitur didapatkan model dengan nilai parameter optimal seperti pada Tabel 2.

Tabel 2. Nilai *Parameter* Optimal Sebelum Melakukan Seleksi Fitur

Nama *Parameter*	Nilai
n_estimators	200
criterion	‘entropy’
max_depth	10
min_samples_split	4
max_features	'log2'

Berdasarkan *parameter* tersebut, dilakukan proses klasifikasi dan didapatkan rata-rata akurasi dan presisi setiap kelas seperti pada Tabel 3.

Tabel 3. Hasil Akurasi Dan Presisi Sebelum Melakukan Seleksi Fitur

Kelas	Rata-Rata Presisi (%)	Rata-Rata Akurasi (%)
Chikungunya	89	88
Dengue	97	
Rift Valley fever	79	
Yellow Fever	100	
Zika	81	
Malaria	83	
Japanese encephalitis	84	
West Nile fever	85	

Setelah melewati tahap seleksi fitur, pencarian model *Random Forest* yang optimal dilakukan sekali lagi menggunakan metode *Random Search*. Hasil pencarian tersebut dapat dilihat pada Tabel 4.

Tabel 4. Nilai *Parameter* Optimal Setelah Melakukan Seleksi Fitur

Nama *Parameter*	Nilai
n_estimators	150
criterion	‘gini’
max_depth	5
min_samples_split	8
max_features	'log2'

Setelah pencarian *parameter* optimal dilakukan, tahap selanjutnya adalah menguji *Random Forest* berdasarkan *parameter* di atas dan fitur yang sudah diseleksi. Adapun hasil dari pengujian tersebut dapat dilihat pada Tabel 5.

Tabel 5. Hasil Akurasi Dan Presisi Setelah Melakukan Seleksi Fitur

Kelas	Rata-Rata Presisi (%)	Rata-Rata Akurasi (%)
Chikungunya	62	76
Dengue	100	
Rift Valley fever	64	
Yellow Fever	71	
Zika	68	
Malaria	81	
Japanese encephalitis	73	
West Nile fever	84	

4. Kesimpulan

Berdasarkan hasil evaluasi yang sudah dilakukan, didapatkan bahwa nilai rata-rata akurasi dari *Random Forest* sebelum dan sesudah melakukan seleksi fitur menggunakan LASSO secara berturut-turut yaitu 88% dan 76%. Dari nilai yang sudah didapat, dapat disimpulkan bahwa *Random Forest* memiliki performa yang lebih baik tanpa adanya seleksi fitur dengan metode LASSO. Hal tersebut dapat terjadi karena beberapa faktor seperti nilai *alpha* dari LASSO yang kurang optimal dalam menentukan fitur-fitur terbaik, adanya fitur yang sebenarnya berpengaruh besar terhadap akurasi dan presisi namun tidak terpilih saat penyeleksian fitur, dan faktor lainnya yang belum bisa penulis dapatkan.

Daftar Pustaka

[1] A. Dinata, *Bersahabat dengan Nyamuk: Jurus Jitu Atasi Penyakit Bersumber Nyamuk*. Arda Publishing House, 2018.

[2] I. Fadilla, P. P. Adikara, dan R. S. Perdana, “Klasifikasi Penyakit Chronic Kidney Disease (CKD) Dengan Menggunakan Metode Extreme Learning Machine (ELM),” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. e-ISSN*, vol. 2548, hlm. 964X, 2018.

[3] R. Wasono, “Perbandingan Metode Random Forest dan naive bayes untuk Klasifikasi Debitur Berdasarkan Kualitas Kredit,” 2022.

[4] M. Parzinger, L. Hanfstaengl, F. Sigg, U. Spindler, U. Wellisch, dan M. Wirnsberger, “Comparison of different training data sets from simulation and experimental measurement with artificial users for occupancy detection — Using machine learning methods Random Forest and LASSO,” *Build Environ*, vol. 223, hlm. 109313, 2022, doi: https://doi.org/10.1016/j.buildenv.2022.109313.

[5] U. Sunarya dan T. Haryanti, “Perbandingan Kinerja Algoritma Optimasi pada Metode Random Forest untuk Deteksi Kegagalan Jantung,” *Jurnal Rekayasa Elektrika*, vol. 18, no. 4, 2022.

[6] V. W. Siburian dan I. E. Mulyana, “Prediksi Harga Ponsel Menggunakan Metode Random Forest,” dalam *Annual Research Seminar (ARS)*, 2019, hlm. 144–147.