Xplore (2023), Vol. 12, No. 1, 36-49
DOI: https://doi.org/10.29244/xplore.v12i1.858

Xplore: Journal of Statistics
Available at https://journal.stats.id/index.php/xplore
e-ISSN: 2655-2744

Perbandingan Performa Metode Pohon Model Logistik dan *Random Forest* pada Pengklasifikasian Data*

Purnama Sari[1], Kusman Sadik[1‡], Mulianto Raharjo[2]

[1]Department of Statistics, IPB University, Indonesia
[2]Kementerian Dalam Negeri Republik Indonesia, Indonesia
[‡]corresponding author: kusmans@apps.ipb.ac.id

Abstract

Multicollinearity and missing data are two common problems in big data. Missing data could decrease the prediction accuracy. Logistic model tree (LMT) is used to handle multicollinearity because multicollinearity does not affect the decision tree. Random forest can be used to decrease variance in prediction case. This study aimed to study the comparison of two methods, LMT and random forest, in multicollinearity and missing data in various cases using simulation study and real data as dataset. Evaluation model is based on classification accuracy and AUC measurement. The result stated that random forest had better performance if the multicollinearity level is moderate. LMT with omitted missing data is proven to have better performance for big data and when a high percentage of missing data occurred, and the multicollinearity level is severe. The next step is analysed real data with different sample size. The result stated that random forest have better performance. Omitted missing data have better performance in classification "breast cancer" data which consist 0,3 % missing data.

Keywords: Dimension of data, LMT, missing data, multicollinearity, random forest.

* Received: Sep 2021; Reviewed: Dec 2022; Published: Jan 2023

1. Pendahuluan

Sebagian besar data yang tersedia tersusun atas peubah-peubah yang tidak saling bebas. Adanya keterkaitan antar peubah penjelas biasanya dapat menimbulkan multikolinearitas. Multikolinearitas merupakan kondisi yang dapat menimbulkan konsekuensi penting pada penafsiran dan penggunaan model regresi linier (Nyrhinen dan Leskinen 2014). Selain peubah-peubah penyusun data yang saling terkait, pada proses pengumpulan data juga tak jarang terdapat *missing data*. Pada kasus klasifikasi *missing data* dapat menjadi sumber umum kelemahan klasifikasi dan salah satu masalah yang dapat memengaruhi hasil prediksi (Malarvizhi dan Thanamani 2012).

Data perlu diolah melalui berbagai proses tertentu agar menjadi sebuah informasi sebagai dasar pengambilan keputusan. Salah satu teknik yang dapat digunakan untuk mengubah data menjadi aturan-aturan keputusan adalah klasifikasi dengan metode pohon keputusan (*decision tree*). *Decision tree* unggul dalam kemampuannya untuk menguraikan proses pengambilan keputusan yang kompleks menjadi lebih sederhana. Hal ini dapat mempermudah dalam interpretasi hasil. Adapun metode klasifikasi yang menerapkan teknik *decision tree* diantaranya adalah pohon model logistik (LMT) dan *random forest*. LMT merupakan metode klasifikasi yang menggabungkan metode pembelajaran pohon keputusan dan regresi logistik pada daun (Chen *et al*. 2017). Dalam konteks data berdimensi rendah, LMT cenderung dianggap sebagai pendekatan standar untuk klasifikasi data biner (Couronne *et al*. 2018). LMT dapat menangani multikolinearitas dengan baik, karena pada metode pohon keputusan adanya multikolinearitas bukanlah suatu masalah (Piramuthu 2008). Pada kasus klasifikasi, *random forest* bekerja dengan k buah pohon dan hasil prediksi akan ditentukan berdasarkan suara terbanyak (*majority vote*). Agregasi sejumlah pohon keputusan yang saling bebas pada *random forest* dapat menghasilkan keragaman yang lebih kecil dibandingkan dengan pohon keputusan tunggal (Couronne *et al*. 2018). Pada perkembangannya, semakin besar dimensi data maka pohon yang terbentuk juga semakin kompleks. Tidak mengikutsertakan *missing data* akan menyebabkan berkurangnya informasi yang bisa diperoleh dalam proses analisis. Oleh karena itu, penanganan dengan memperkirakan nilai *missing data* dapat menjadi alternatif yang mungkin lebih baik. Penanganan missing data dapat dilakukan dengan metode imputasi, salah satunya dengan menggunakan nilai *mean*.

Berdasarkan uraian tersebut, dalam penelitian ini dilakukan kajian perbandingan performa metode LMT dan *random forest* pada pengklasifikasian data dengan dimensi data yang berbeda-beda. Persentase *missing data* dan tingkat multikolinearitas pada kasus klasifikasi juga menjadi hal yang dikaji. Kajian dilakukan pada data simulasi dan data riil. Data simulasi merupakan beberapa data hasil pembangkitan bilangan acak dengan skenario yang dibuat untuk menjawab tujuan penelitian. Data riil yang digunakan adalah empat data yang diambil dari UCI *machine learning repository*. Adapun untuk keempat data mempunyai kategori kelas data yang tidak seimbang. Ketidakseimbangan kelas pada kasus klasifikasi dapat menyebabkan kecenderungan prediksi ke kelas mayoritas sehingga dilakukan penanganan dengan metode SMOTE (Baizal *et al*. 2009).

2. Metodologi

2.1 Data

Perbandingan performa metode LMT dan *random forest* dilakukan pada dua jenis data, yaitu data simulasi dan data riil.

2.1.1 Data Simulasi

Kajian simulasi dilakukan untuk mengevaluasi performa LMT dan *random forest* pada data yang mengandung *missing data*, dan terdapat tingkatan multikolinearitas sedang dan tinggi. Kajian simulasi melibatkan beberapa data hasil pembangkitan yang dibuat dengan tingkat multikolinearitas dan persentase *missing data* yang berbeda, serta banyak amatan yang berbeda pula. Banyak amatan dibuat dua taraf yaitu 300 dan 10.000 amatan. Penentuan banyaknya amatan disesuaikan dengan kebutuhan penelitian dan dengan mempertimbangkan kemampuan *software* yang digunakan. Selanjutnya tingkat multikolinearitas, dibuat dua taraf yaitu multikolinearitas sedang dengan koefisien korelasi berkisar antara 0,50 sampai 0,70 dan multikolinearitas tinggi dengan koefisien korelasi berkisar antara 0,97 sampai 0,99 (Sungkono dan Nugrahaningsih 2017). Selanjutnya untuk persentase *missing data* juga dibuat dua taraf, yaitu 1% dan 8% dari total banyaknya satuan amatan dalam data. Proporsi *missing data* berkaitan langsung dengan kualitas inferensi statistik, namun belum ada batasan yang tegas terkait persentase *missing data* yang dapat diterima dalam sebuah data untuk menghasilkan inferensi statistik yang valid (Bunawolo 2017). Penelitian Schafer (1999) dalam Bunawolo (2017) menegaskan bahwa tingkat *missing data* kurang dari 5% tidak memberikan pengaruh. Oleh karena itu, dalam penelitian ini *missing data* dibuat dengan persentase kurang dari dan lebih dari 5%. *Missing data* dibuat secara acak keseluruhan atau *missing completely at random* (MCAR) pada setiap baris dan kolom. Terdapat total 8 skenario simulasi yang dibuat dengan banyaknya peubah penjelas adalah 10 untuk masing-masing skenario.

2.1.2 Data Riil

Analisis data riil dilakukan untuk mengevaluasi performa LMT dan *random forest* pada data dengan dimensi data yang berbeda-beda dan dengan berbagai tingkat persentase *missing data*. Data riil yang digunakan merupakan empat data yang diambil dari https://archive.ics.uci.edu/ml/index.php dengan spesifikasi sebagai berikut.

Tabel 1: Spesifikasi data riil

Nama data	Banyaknya amatan	Banyaknya peubah
Adult	48.842	15
Breast Cancer	286	10
Hepatitis	155	20
Bank Marketing	41.188	21

2.2 Prosedur Analisis Data

2.2.1 Data Simulasi

Pembangkitan data simulasi dilakukan menggunakan *software R version 4.0.4* dengan memanfaatkan beberapa *package*, diantaranya “MASS”, “DMwR”, “randomForest”, “RWeka”, dan “ROCR”. Prosedur simulasi yang dilakukan adalah sebagai berikut:

1. Menentukan 8 skenario simulasi yang akan dibangkitkan, disajikan pada tabel berikut.

Tabel 2: Skenario data simulasi

Skenario	Tingkat multikolinearitas	Persentase *missing data* (%)	Banyaknya amatan
Skenario 1	Sedang	1	300
Skenario 2	Sedang	8	300
Skenario 3	Sedang	1	10.000
Skenario 4	Sedang	8	10.000
Skenario 5	Tinggi	1	300
Skenario 6	Tinggi	8	300
Skenario 7	Tinggi	1	10.000
Skenario 8	Tinggi	8	10.000

2. Menentukan matriks korelasi yang digunakan untuk membangkitkan peubah penjelas. Selanjutnya dilakukan pembangkitan 10 peubah penjelas sebanyak n amatan dengan fungsi "*mvrnorm*" dari *package* "MASS" dalam RStudio.
3. Membangkitkan n bilangan acak sisaan menyebar normal $\varepsilon \sim \mathrm{N}(0{,}20)$, kemudian bangkitkan nilai y dari model regresi linier berganda sebagai berikut.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{y} = \begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_n} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x_{11}} & \dots & \boldsymbol{x_{1p}} \\ \mathbf{1} & \boldsymbol{x_{21}} & \dots & \boldsymbol{x_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{1} & \boldsymbol{x_{n1}} & \dots & \boldsymbol{x_{np}} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \vdots \\ \boldsymbol{\beta_p} \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_n} \end{bmatrix}$$

4. Selanjutnya melakukan transformasi nilai y dengan ketentuan sebagai berikut:

$$y = \begin{cases} \{1, \text{untuk } y_i \geq \ \mathrm{Q3} \\ \{0, \text{untuk } y_i < \ \mathrm{Q3} \end{cases}$$

5. Melakukan *sampling* indeks baris dan kolom untuk dibuat *missing data*.
6. Selanjutnya dilakukan penanganan *missing data* dengan dua metode, pertama dengan menghapus *missing data* dan kedua melakukan imputasi *missing data* dengan nilai *mean* masing-masing kolom.
7. Melakukan pembagian data menjadi data latih dan data uji dengan rasio $7:3$ untuk keperluan tahap pemodelan.
8. Selanjutnya dilakukan penanganan ketidakseimbangan kelas pada data latih dengan teknik SMOTE yang memanfaatkan *package* "DMwR" pada RStudio.
9. Membangun model dengan metode *random forest* dan LMT. Parameter di setiap model menggunakan *default parameters* masing-masing model. Pemodelan *random forest* dimulai dari tahapan *bootstrap*, pembentukan pohon klasifikasi secara rekursif. Setiap pohon kemudian dilakukan penyekatan dengan memilih peubah pemilah terbaik berdasarkan nilai *information gain* yang dirumuskan sebagai berikut:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy\ (i) \right)$$

yaitu memilih peubah dengan nilai *information gain* terbesar. Adapun untuk model LMT pohon keputusan dibentuk dengan algoritma *LogitBoost*, dan terakhir dilakukan pemangkasan pohon.

10. Melakukan prediksi kelas pada data uji. Prediksi final pada *random forest* merupakan suara terbanyak dari hasil prediksi k buah pohon. Adapun untuk hasil dari pemodelan dengan LMT merupakan nilai peluang untuk masing-

masing kelas, sehingga untuk memperoleh kelas amatan ditentukan dengan menetapkan nilai *cut off* tertentu.

11. Ulangi langkah 2 sampai 10 sebanyak 100 kali untuk masing-masing skenario.
12. Kemudian dilakukan perhitungan rataan nilai akurasi, sensitivitas, spesifisitas, dan skor F1 masing-masing model dalam 100 kali ulangan.
13. Menghitung luas *area under curve* (AUC) dari kurva *receiver operator characteristic* (ROC) masing-masing model dan membandingkannya.

2.2.2 Data Riil

Prosedur analisis data riil yang dilakukan adalah sebagai berikut:

1. Melakukan eksplorasi data dengan membuat *bar plot* peubah respon untuk mengetahui perbandingan kategori kelas data, kemudian dibuat plot perbandingan persentase *missing data* yang ada pada setiap set data.
2. Melakukan pra analisis data yaitu melakukan *cleaning data* dengan mengubah beberapa karakter tidak terdeteksi pada data seperti tanda "?" dan label "*unknown*" dimasukkan ke dalam kategori *missing data*.
3. Melakukan penanganan *missing data* dengan dua metode, pertama dengan menghapus *missing data* dan kedua melakukan imputasi *missing data*.
4. Tahapan berikutnya adalah sama dengan prosedur simulasi poin 7 sampai 10, dengan 10 kali ulangan.

3. Hasil dan Pembahasan

3.1 Kajian Simulasi

3.1.1 Deskripsi Data Hasil Bangkitan

Data hasil bangkitan terdiri atas delapan set data, empat set data memiliki persentase *missing data* sebanyak 1% dan yang lainnya 8%.

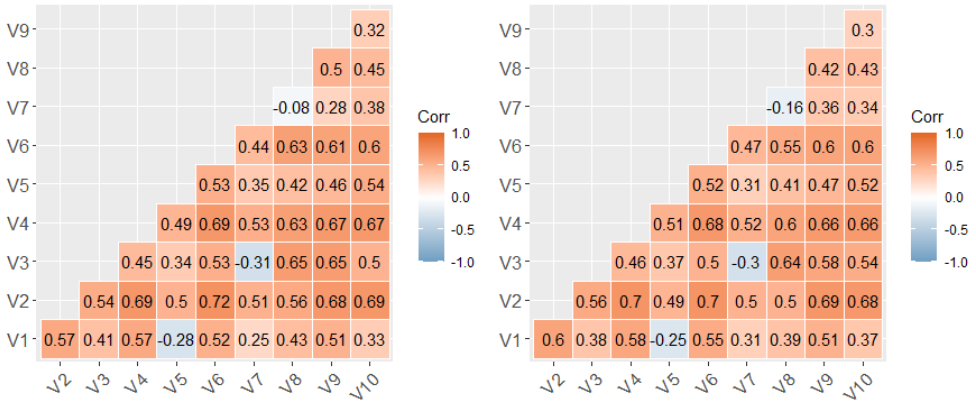
Tabel 3: Persentase *missing data* di setiap peubah data simulasi

Skenario simulasi	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1 dan 5	0,33	0,33	1,00	0,67	0,33	1,33	1,00	1,00	2,00	2,00
2 dan 6	10,00	6,33	10,30	6,67	9,00	7,67	8,33	6,33	6,33	9,00
3 dan 7	1,19	0,93	0,93	0,92	1,18	0,92	1,04	0,85	0,99	1,05
4 dan 8	7,96	7,58	8,26	7,74	8,35	7,78	8,09	7,59	8,01	8,64

Skenario *missing data* yang dibuat terdistribusi secara acak untuk seluruh unit amatan, artinya adanya *missing data* tidak berkaitan dengan nilai semua peubah. Indeks baris dan kolom untuk missing data ditentukan secara acak dengan melakukan sampling terhadap indeks baris dan kolom dari data. Sehingga setiap peubah penjelas mempunyai *missing data* dengan persentase yang beragam seperti pada Tabel 3. Keberadaan *missing data* yang cukup besar dapat memengaruhi hasil analisis lanjutan. Oleh karena itu, perlu dilakukan penanganan terhadap keberadaan *missing data*.

Peubah penjelas dibangitkan dengan menggunakan fungsi *multivariate normal* "*mvrnorm*" pada *package* "MASS" dengan menetapkan nilai *mean* adalah 1. Peubah

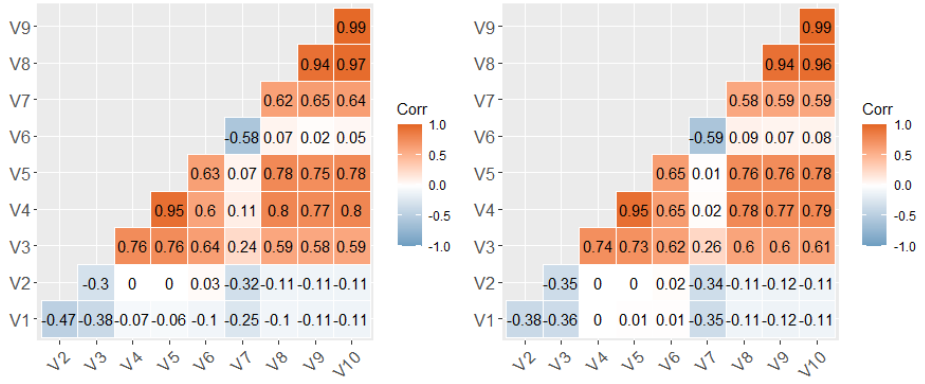
penjelas yang dibangkitkan beberapa diantaranya mempunyai hubungan linier yang cukup kuat terlihat pada plot korelasi.



(a) Skenario 1&2 (b) Skenario 3&4

Gambar 1: Plot korelasi antar peubah penjelas

Gambar 1 merupakan visualisasi dari nilai koefisien korelasi antar peubah penjelas untuk skenario multikolinearitas dengan korelasi sedang Nilai korelasi yang berkisar antara 0,50 sampai 0,70 seperti di atas dapat mengindikasikan adanya multikolinearitas sedang (Sungkono dan Nugrahaningsih 2017). Adapun untuk skenario 5, 6, 7, dan 8 dibuat beberapa peubah dengan koefisien korelasi yang tinggi seperti berikut ini.



(a) Skenario 5&6 (b) Skenario 7&8

Gambar 2: Plot korelasi antar peubah penjelas

Nilai korelasi yang mencapai 0,99 dapat mengindikasikan adanya multikolinearitas tinggi (Sungkono dan Nugrahaningsih 2017).

3.1.2 Pemodelan dan Evaluasi Model Klasifikasi

Proses pemodelan dilakukan dengan pengulangan sebanyak 100 kali, kemudian dihitung rataan dari setiap hasil evaluasi kinerja dari tabel ketepatan klasifikasi. Berikut disajikan tabel perbandingan nilai akurasi, sensitivitas, spesifisitas, dan skor F1 dari masing-masing model.

Tabel 4: Nilai ketepatan klasifikasi pemodelan data skenario 1

Ukuran evaluasi	*Random forest*		LMT	
	NA dihapus	Imputasi NA	NA dihapus	Imputasi NA
Akurasi	0,6914	0,7000	0,7654*	0,5556
Sensitivitas	0,7049	0,7042	0,7541*	0,5211
Spesifisitas	0,6500	0,6842	0,8000*	0,6842
Skor F1	0,7748	0,7874	0,8288*	0,6491

*kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 terbaik

Tabel 4 menunjukkan untuk skenario 1 dengan *missing data* 1% dan banyaknya amatan 300 penerapan LMT dengan penghapusan *missing data* menghasilkan kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 yang lebih baik dibandingkan tiga model lainnya. Nilai yang dihasilkan berturut-turut adalah 0,7654; 0,7541; 0,8000; dan 0,8288.

Tabel 5: Nilai ketepatan klasifikasi pemodelan data skenario 2

Ukuran evaluasi	*Random forest*		LMT	
	NA dihapus	Imputasi NA	NA dihapus	Imputasi NA
Akurasi	0,6667*	0,6556	0,6111	0,6111
Sensitivitas	0,6667*	0,7286	0,6000	0,6714
Spesifisitas	0,6667*	0,4000	0,6667	0,4000
Skor F1	0,7692*	0,7669	0,7200	0,7287

*kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 terbaik

Skenario 2 dengan *missing data* 8% dan banyaknya amatan 300, penerapan *random forest* dengan penghapusan *missing data* menghasilkan kombinasi nilai akurasi sebesar 0,6667; sensitivitas sebesar 0,6667; spesifisitas sebesar 0,6667; dan skor F1 sebesar 0,7692 yang dinilai lebih baik dibandingkan dengan tiga model lainnya seperti disajikan pada Tabel 5.

Tabel 6: Nilai ketepatan klasifikasi pemodelan data skenario 3

Ukuran evaluasi	*Random forest*		LMT	
	NA dihapus	Imputasi NA	NA dihapus	Imputasi NA
Akurasi	0,6556	0,6667*	0,6211	0,6213
Sensitivitas	0,6681	0,6742*	0,6115	0,5918
Spesifisitas	0,6173	0,6442*	0,6502	0,7090
Skor F1	0,7448	0,7516*	0,7083	0,7004

*kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 terbaik

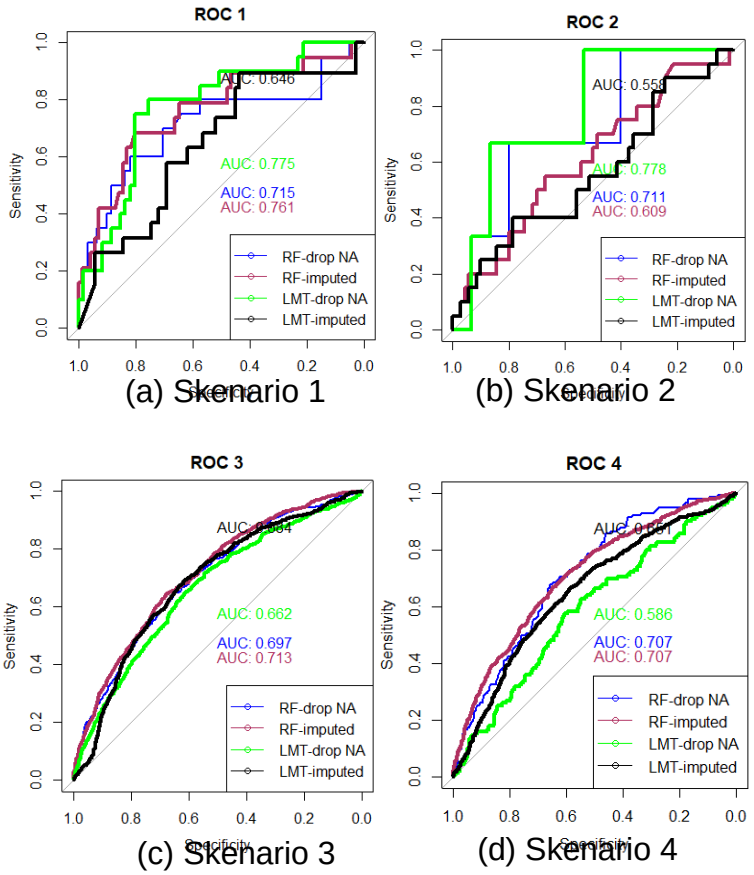
Skenario 3 dengan *missing data* 1% dan 10.000 amatan, penerapan *random forest* dengan imputasi *missing data* menghasilkan kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 yang lebih baik. Adapun nilai yang dihasilkan berdasarkan Tabel 6 berturut-turut adalah sebesar 0,6556; 0,6681; 0,6173; dan 0,7448.

Tabel 7: Nilai ketepatan klasifikasi pemodelan data skenario 4

Ukuran evaluasi	*Random forest*		LMT	
	NA dihapus	Imputasi NA	NA dihapus	Imputasi NA
Akurasi	0,6683	0,6543*	0,6167	0,6133
Sensitivitas	0,6833	0,6527*	0,6594	0,6055
Spesifisitas	0,6187	0,6589*	0,4748	0,6358
Skor F1	0,7600	0,7367*	0,7255	0,6989

*kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 terbaik

Berikutnya untuk skenario 4 memiliki *missing data* 8% dan banyak amatan 10.000, penerapan *random forest* dengan imputasi *missing data* menghasilkan kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 yang lebih baik dibandingkan dengan tiga model lainnya. Tabel 7 menunjukkan nilai yang dihasilkan berturut-turut adalah sebesar 0,6543; 0,6527; 0,6589; dan 0,7367.



(a) Skenario 1 (b) Skenario 2

(c) Skenario 3 (d) Skenario 4

Gambar 3: Perbandingan luas AUC model LMT dan *random forest*

Berdasarkan uraian hasil evaluasi model dengan mengacu pada nilai akurasi, sensitivitas, spesifisitas, skor F1, dan nilai AUC di atas, untuk data dengan tingkat multikolinearitas sedang, *random forest* dengan imputasi *missing data* memiliki performa yang lebih baik pada data dengan banyaknya amatan dan persentase *missing data* yang besar (10.000 amatan dan 8% NA).

Tabel 8: Nilai ketepatan klasifikasi pemodelan data skenario 5

Ukuran evaluasi	*Random forest*		LMT	
	NA dihapus	Imputasi NA	NA dihapus	Imputasi NA
Akurasi	0,6420	0,6444*	0,5185	0,5889
Sensitivitas	0,6393	0,6970*	0,5082	0,6061
Spesifisitas	0,6500	0,5000*	0,5500	0,5417
Skor F1	0,7290	0,7419*	0,6139	0,6838

*kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 terbaik

Berikutnya adalah nilai akurasi, sensitivitas, spesifisitas, dan skor F1 untuk skenario 5, 6, 7, dan 8 yang memiliki tingkat multikolinearitas tinggi. Skenario 5 dengan *missing data* 1% dan banyaknya amatan 300, penerapan *random forest* dengan imputasi *missing data* menghasilkan kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 yang lebih baik dibandingkan dengan tiga model lainnya. Tabel 8 menunjukkan nilai yang dihasilkan berturut-turut adalah sebesar 0,6444; 0,6970; 0,5000; dan 0,7419. Evaluasi model juga dilakukan dengan mengevaluasi nilai AUC dari kurva ROC. Hasil yang diperoleh dapat dilihat pada Gambar 4. Model *random forest* dengan imputasi *missing data* pada skenario 5 menghasilkan nilai AUC terbesar yaitu sebesar 0,665 dapat dilihat pada Gambar 4(a).

Tabel 9: Nilai ketepatan klasifikasi pemodelan data skenario 6

Ukuran evaluasi	*Random forest*		LMT	
	NA dihapus	Imputasi NA	NA dihapus	Imputasi NA
Akurasi	0,8889*	0,5222	0,7778	0,4556
Sensitivitas	0,8667*	0,5152	0,8667	0,4697
Spesifisitas	1,0000*	0,5417	0,3333	0,4167
Skor F1	0,9286*	0,6126	0,8667	0,5586

*kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 terbaik

Skenario 6 dengan *missing data* 8% dan banyaknya amatan 300, penerapan *random forest* dengan menghapus *missing data* menghasilkan kombinasi nilai akurasi sebesar 0,8889; sensitivitas yaitu 0,8667; spesifisitas sebesar 1,0000; dan skor F1 sebesar 0,9286 seperti pada Tabel 9 yang dinilai lebih baik dibandingkan tiga model lainnya. Nilai AUC untuk model ini adalah sebesar 0,867 dapat dilihat pada Gambar 4(b).

Tabel 10: Nilai ketepatan klasifikasi pemodelan data skenario 7

Ukuran evaluasi	*Random forest*		LMT	
	NA dihapus	Imputasi NA	NA dihapus	Imputasi NA
Akurasi	0,6248	0,6233	0,6393	0,6417*
Sensitivitas	0,6438	0,6234	0,6447	0,6340*
Spesifisitas	0,5665	0,6231	0,6224	0,6653*
Skor F1	0,7215	0,7142	0,7296	0,7276*

*kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 terbaik

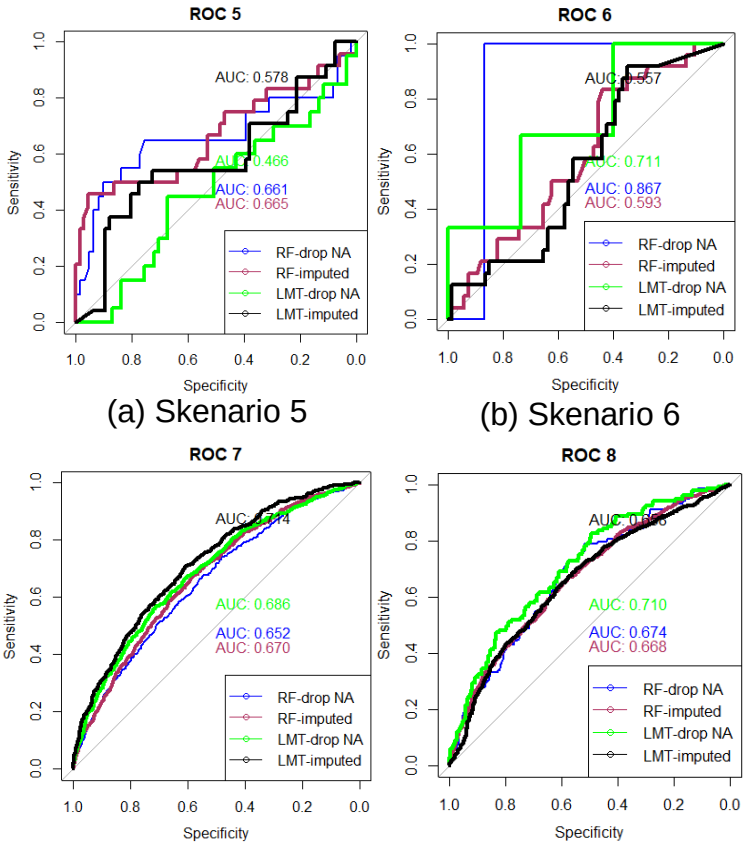
Berikutnya skenario 7 dengan *missing data* 1% dan 10.000 amatan, penerapan LMT dengan imputasi *missing data* menghasilkan kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 yang lebih baik dibandingkan dengan tiga model lainnya. Tabel 10 menunjukkan nilai yang dihasilkan berturut-turut adalah sebesar 0,6417; 0,6340; 0,6653; dan 0,7276. Adapun nilai AUC yang peroleh dari penerapan model ini adalah sebesar 0,714 seperti pada Gambar 4(c).

Tabel 11: Nilai ketepatan klasifikasi pemodelan data skenario 8

Ukuran evaluasi	*Random forest*		LMT	
	NA dihapus	Imputasi NA	NA dihapus	Imputasi NA
Akurasi	0,6383	0,6273	0,6400*	0,6327
Sensitivitas	0,6531	0,6366	0,6467*	0,6504
Spesifisitas	0,5865	0,5995	0,6165*	0,5794
Skor F1	0,7376	0,7194	0,7366*	0,7266

*kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 terbaik

Terakhir untuk skenario 8 dengan *missing data* 8% dan banyaknya amatan 10.000, penerapan LMT dengan penghapusan *missing data* menghasilkan kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 yang lebih baik dibandingkan dengan tiga model lainnya. Tabel 11 menunjukkan nilai yang dihasilkan berturut-turut adalah sebesar 0,6400; 0,6467; 0,6165 dan 0,7366. Adapun nilai AUC yang dihasilkan adalah sebesar 0,71 seperti pada Gambar 4(d).



(a) Skenario 5 (b) Skenario 6

(c) Skenario 7 (d) Skenario 8

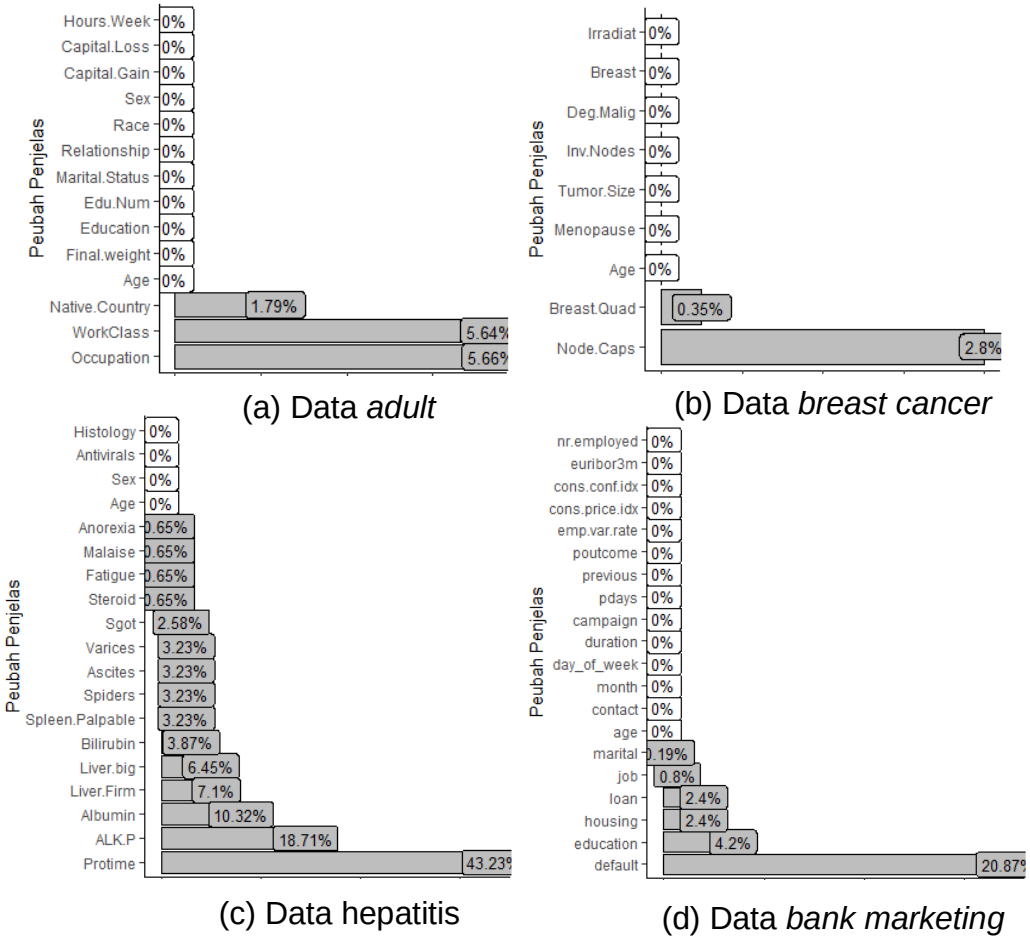
Gambar 4: Perbandingan luas AUC model LMT dan *random forest*

Berdasarkan uraian hasil evaluasi model dengan mengacu pada nilai akurasi, sensitivitas, spesifisitas, skor F1, dan nilai AUC di atas, untuk data dengan tingkat multikolinearitas tinggi, banyaknya amatan dan persentase *missing data* besar (10.000 amatan dan 8% NA), LMT dengan penghapusan *missing data* terbukti menghasilkan performa yang lebih baik.

3.2 Analisis Data Riil

3.2.1 Eksplorasi data

Data yang digunakan dalam penelitian ini diambil dari UCI *machine learning repository*. Keempat data yang digunakan merupakan data dengan kategori kelas yang tidak seimbang. Masing-masing data memiliki *missing data* dengan persentase yang berbeda-beda, seperti disajikan pada Gambar 5. Data hepatitis mempunyai persentase data hilang yang terbesar dibandingkan 3 data lainnya mencapai 5,4%.



(a) Data *adult* (b) Data *breast cancer*

(c) Data hepatitis (d) Data *bank marketing*

Gambar 5: Persentase nilai *missing data* pada data riil

3.2.2 Pemodelan dan Evaluasi Model Klasifikasi

Sama seperti kajian simulasi, pemodelan klasifikasi untuk data riil juga dilakukan dengan dua metode yaitu *random forest* dan LMT dengan dua jenis penanganan terhadap *missing data* yaitu penghapusan dan imputasi *missing data*. Imputasi dilakukan dengan menggunakan nilai *mean* untuk data numerik, dan nilai modus untuk data kategorik. Hasil pemodelan dievaluasi menggunakan ukuran ketepatan klasifikasi dan nilai AUC yang merupakan luas dari kurva ROC. Proses pemodelan

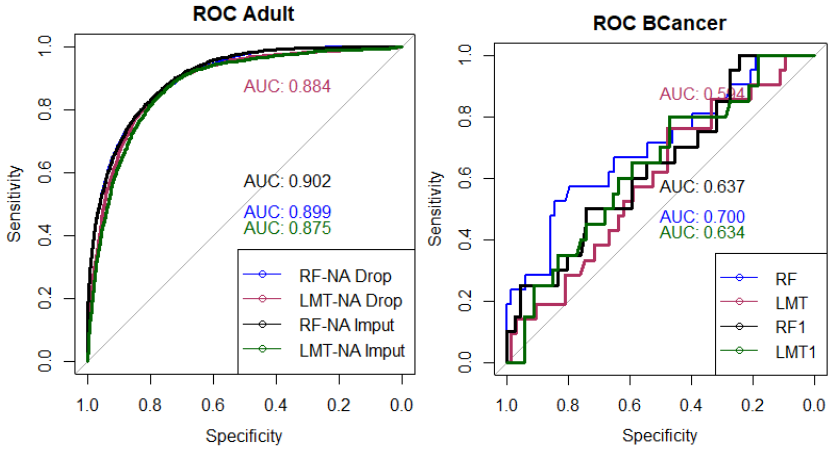
dilakukan dengan pengulangan sebanyak 10 kali, kemudian dihitung rataan dari setiap hasil evaluasi kinerja dari tabel ketepatan klasifikasi. Berikut disajikan tabel perbandingan nilai akurasi, sensitivitas, spesifisitas, dan skor F1 dari masing-masing model.

Tabel 12: Nilai akurasi, sensitivitas, spesifisitas, dan skor F1

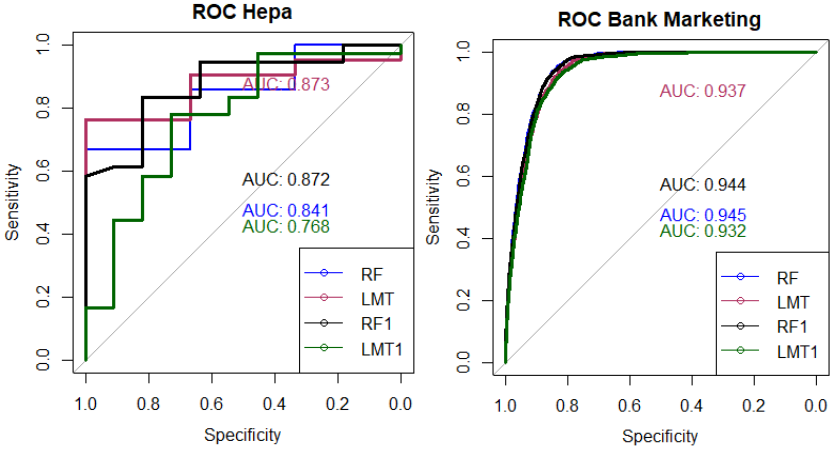
Data	Ukuran evaluasi	*Random forest*		LMT	
		NA dihapus	Imputasi NA	NA dihapus	Imputasi NA
Adult	Akurasi	0,8250	0,8259*	0,8238	0,8252
	Sensitivitas	0,8397	0,8390*	0,8558	0,8511
	Spesifisitas	0,7812	0,7849*	0,7286	0,7435
	Skor F1	0,8778	0,8798*	0,8790	0,8808
Breast Cancer	Akurasi	0,6500*	0,6523	0,6226	0,6105
	Sensitivitas	0,6597*	0,7113	0,6426	0,6700
	Spesifisitas	0,6238*	0,5081	0,5666	0,4607
	Skor F1	0,7298*	0,7416	0,7086	0,7048
Hepatitis	Akurasi	0,8042	0,7936*	0,7917	0,7787
	Sensitivitas	0,5821	0,7585*	0,5821	0,6929
	Spesifisitas	0,8329	0,8009*	0,8185	0,8034
	Skor F1	0,5465	0,6050*	0,4839	0,5643
Bank Marketing	Akurasi	0,8722	0,8776*	0,8650	0,8718
	Sensitivitas	0,8720	0,8772*	0,8672	0,8744
	Spesifisitas	0,8741	0,8810*	0,8500	0,8513
	Skor F1	0,9225	0,9271*	0,9180	0,9237

*kombinasi nilai akurasi, sensitivitas, spesifisitas, dan skor F1 terbaik

Tabel 12 menunjukkan pemodelan *random forest* dengan imputasi *missing data* menghasilkan performa yang lebih baik dibandingkan LMT untuk data "*adult", "*hepatitis", dan "*bank marketing*". Adapun untuk data "*breast cancer*", penghapusan *missing data* terbukti menghasilkan performa yang lebih baik. Evaluasi model juga dilakukan dengan mengevaluasi kurva ROC. ROC dinilai lebih baik untuk mengevaluasi kinerja model klasifikasi data tidak seimbang karena merupakan plot antara nilai sensitivitas dan nilai spesifisitas. Hasil yang diperoleh dapat dilihat pada Gambar 6.



(a) Data *adult* (b) Data *breast cancer*



(c) Data hepatitis (d) Data *bank marketing*

Gambar 6: Perbandingan luas AUC model LMT dan RF data riil

Berdasarkan uraian hasil evaluasi model dengan mengacu pada nilai akurasi, sensitivitas, spesifisitas, skor F1, dan nilai AUC di atas, untuk keempat data yang digunakan pemodelan menggunakan *random forest* menghasilkan performa yang lebih baik. Adapun untuk data "*breast cancer*" yang memiliki persentase *missing data* hanya 0,3% penghapusan *missing data* menghasilkan performa yang lebih baik dengan nilai AUC yang dihasilkan adalah sebesar 0,7. Pemodelan dengan *random forest* menghasilkan performa yang lebih baik mungkin berkaitan dengan banyaknya peubah penjelas di dalam data, mengingat *random forest* dapat mengevaluasi kepentingan peubah dalam model.

4. Simpulan

Kajian simulasi yang dilakukan menunjukkan penerapan metode random forest dengan imputasi missing data menggunakan nilai mean menghasilkan performa yang lebih baik untuk memodelkan data dengan tingkat multikolinearitas sedang (koefisien korelasi berkisar antara 0,50 sampai 0,70), banyaknya amatan, dan persentase missing data yang besar (10.000 amatan, 10 peubah penjelas, dan 8% missing data). Sedangkan untuk data dengan tingkat multikolinearitas tinggi (koefisien korelasi berkisar antara 0,97 sampai 0,99), banyaknya amatan, dan persentase missing data yang besar (10.000 amatan, 10 peubah penjelas, dan 8% missing data), penerapan

metode LMT dengan penghapusan missing data terbukti menghasilkan performa yang lebih baik dibandingkan imputasi missing data menggunakan nilai mean. Sehingga, metode imputasi dengan nilai mean belum menunjukkan hasil yang lebih unggul dibandingkan dengan penghapusan missing data ketika diterapkan pada metode LMT.

Selanjutnya, analisis data riil dilakukan dengan memodelkan empat data yang memiliki dimensi data yang berbeda. Hasil yang diperoleh menunjukkan penerapan random forest dengan imputasi missing data menggunakan nilai mean menghasilkan performa yang lebih baik untuk data “adult”, “hepatitis”, dan “bank marketing”. Adapun untuk data “breast cancer” yang memiliki persentase missing data hanya 0,30%, penghapusan missing data menghasilkan performa yang lebih baik. Hal ini mungkin berkaitan dengan kemampuan random forest dapat melakukan pemilihan peubah penjelas secara acak yang akan digunakan untuk pemodelan dan dapat mengekstrak tingkat kepentingan peubah dalam model.

Daftar Pustaka

Baizal ZA, Bijaksana MA, Sastrawan AS. 2009. Analisis pengaruh metode over sampling dalam churn prediction untuk perusahaan telekomunikasi. Seminar Nasional Aplikasi Teknologi Informasi 2009 (SNATI 2009). ISSN: 1907-5022.

Bunawolo EN. 2017. Imputasi Data Hilang pada Survei Industri Besar Sedang Sumatera Utara Menggunakan Fuzzy C-Means Dioptimalkan dengan Algoritma Genetika [Tesis]. Surabaya (ID): Institut Teknologi Sepuluh Nopember.

Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Duan Z, Ma J. 2017. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena. 151:147-160.

Couronne R, Probst P, Laure B. 2018. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 19: 270.

Malarvizhi R, Thanamani AS. 2012. K-NN classifier performs better than k-means clustering in missing value imputation. Journal of Computer Engineering. 6(5):12-15.

Nyrhinen JN, Leskinen E. 2014. Multicolinearity in marketing models: notes on the application of ridge trace estimation in structural equation modeling. Electronic Journal of Business Research Methods. 12(1): 3-15.

Piramuthu S. 2008. Input data for decision trees. Expert System with Application. 34: 1220-1226.

Sungkono J, Nugrahaningsih TK. 2017. Simulasi dampak multikolinearitas pada kondisi penyimpangan asumsi normalitas. Magistra, (101): 45-50.