INNOVATIVE: Journal Of Social Science Research
Volume 4 Nomor 1 Tahun 2024 Page 11773-11783
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Meningkatkan Kinerja Model Klasifikasi Curah Hujan Melalui Penanggulangan *Missing Value* Dengan Imputasi Berbasis Model

Winalia Agwil[1✉], Dian Agustina[2], Herlin Fransiska[3], Idam Abdurrohim Hasani [4]
Dosen S1 Statistika, Universitas Bengkulu
Email: winaliaagwil@unib.ac.id[1✉]

Abstrak

Penanggulangan terhadap data yang tidak lengkap telah banyak dikembangkan. Salah satu metode yang digunakan untuk menangani data hilang (*missing value*) adalah imputasi. Beberapa metode imputasi berbasis model untuk menduga nilai hilang pada variabel numerik adalah menggunakan metode regresi dan *Random Forest*. Penerapan metode imputasi ini dilakukan pada data curah hujan di Provinsi Bengkulu dari tahun 2013 sampai 2022 dengan variabel yang digunakan Y1 = Curah Hujan, X1 = Suhu Rata-rata, X2 = Suhu Maksimum, X3 = Suhu Minimum, X4 = Tekanan Udara diatas Permukaan Laut, X5 = Arah Angin, X6 = Kecepatan Angin Maksimum, X7 = Tingkat Awan, X8 = Lama Penyinaran Matahari. Hasil diperoleh bahwa kedua metode imputasi berbasis model yang digunakan memberikan nilai akurasi yang tidak terlalu berbeda. Nilai akurasi konsisten antara data training dan data testing, hal ini mengindikasikan bahwa imputasi yang dilakukan sudah baik. Pemodelan dilakukan dengan menggunakan metode CART dengan variabel dengan kontribusi tinggi adalah variabel yang memiliki kontribusi paling besar adalah arah angin pada hari sebelumnya dan variabel tutupan awan pada satu hari sebelumnya.

Kata Kunci : *CART, Imputasi Regresi, missing value, CART*

Abstract

There are many ways to deal with incomplete data (missing value). One of the methods used to deal with lost data is imputation. Several model-based imputation methods to estimate missing values on numerical variables are using regression method and Random Forest. The application of this imputation method was carried out on rainfall data in Bengkulu Province from 2013 to 2022 with the variables used Y1 = Rainfall, X1 = Average Temperature, X2 = Maximum Temperature, X3 = Minimum Temperature, X4 = Air Pressure above Sea Level, X5 = Wind Direction, X6 = Maximum Wind Speed, X7 = Cloud Level, X8 = Sunshine Duration. The results showed that the model-based imputation methods used provide not too different accuracy values. The accuracy value is consistent between the training data and the testing data, this indicates that the imputation performed is good. The modeling is carried out using the CART method with the variable with the highest contribution being the variable that has the greatest contribution, namely the wind direction on the previous day and the cloud cover variable on the previous day

Keyword: *CART, missing value, random forest impute, regression impute*

PENDAHULUAN

Kenaikan suhu bumi sangat berdampak terhadap perubahan sistem iklim yang mempengaruhi berbagai aspek pada perubahan alam dan kehidupan manusia. Beberapa contoh dampak negatif perubahan iklim adalah gagal panen, cuaca ekstrim, dan meningkatnya wabah penyakit. Salah satu faktor penentu yang mempengaruhi tipe iklim adalah curah hujan. Indonesia dikenal dengan negara yang memiliki iklim tropis dengan dua musim yaitu, musim kemarau dan musim hujan. Musim hujan adalah musim dengan ciri meningkatnya curah hujan di suatu wilayah dibandingkan biasanya dalam jangka waktu tertentu secara tetap. Ada beberapa dampak yang ditimbulkan oleh curah hujan tinggi yakni terganggunya sistem logistik nasional, keselamatan serta mobilisasi masyarakat pengguna jalan akan terganggu, serta kerusakan infrastruktur karena kerusakan jalan akibat longsor dan banjir. Diperlukan model atau sistem yang dapat mengklasifikasikan curah hujan dengan akurasi tinggi sehingga efek negatifnya dapat dicegah dengan tindakan preventif.

Pengklasifikasian curah hujan dapat dilakukan menggunakan metode machine learning (ML) dengan memanfaatkan historis data iklim lainnya seperti suhu, kelembaban udara, kecepatan angin, arah angin dan lama penyinaran matahari. Salah satu metode klasifikasi dalam ML yang dapat digunakan adalah Random Forest. Model klasifikasi yang dibentuk dengan metode random forest dipercaya memiliki akurasi yang tinggi, namun hal ini tidak akan tercapai dengan baik apabila di dalam dataset terdapat permasalahan seperti data yang tidak lengkap. Data curah hujan dan iklim lainnya yang digunakan pada penelitian ini merupakan data harian dari tahun 2016 sampai tahun 2020 yang diperoleh dari Badan

Meteorologi, Klimatologi, dan Geofisika (BMKG). Dewasanya, pada record data BMKG seringkali ditemui data yang tidak lengkap yang biasanya disebut nilai hilang (missing value). Situasi ini dapat mempengaruhi proses ML karena sebagian besar algoritma ML tidak dapat langsung diterapkan pada data yang tidak lengkap. Selain itu, ketidaklengkapan data akan mempengaruhi keakuratan hasil klasifikasi (Acuna dan Rodriguez, 2004). Sehingga, diperlukan algoritma yang tepat dalam menangani nilai hilang agar penggunaan algoritma klasifikasi menjadi lebih efektif dan mendukung ketepatan pengklasifikasian curah hujan.

Penanggulangan terhadap data yang tidak lengkap telah banyak dikembangkan. Secara umum, penanggulangannya dapat dikategorikan menjadi dua, yaitu metode imputasi dan amputasi. Dalam beberapa kondisi, seringkali metode amputasi tidak disarankan untuk digunakan dan lebih efektif jika menggunakan metode imputasi. Beberapa penelitian juga menunjukkan bahwa penggunaan imputasi untuk menangani data yang hilang atau tidak lengkap dapat meningkatkan akurasi klasifikasi dibandingkan dengan kasus dimana imputasi tidak digunakan (Laencina dkk., 2009). Imputasi data yang tidak lengkap bertujuan untuk memberikan perkiraan nilai yang hilang dengan mempelajari karakteristik data yang diamati (Batista dan Monard, 2003).

Beberapa peneliti telah menerapkan metode imputasi dalam menangani missing value. Fadillah dan Puspita (2020) telah melakukan penanggulangan nilai hilang menggunakan metode imputasi KNN terboboti, hasil penelitian menunjukkan bahwa metode ini efektif dalam menanggulangi nilai hilang pada data indeks produksi Industri Mikro kecil (IMK). Penelitian lainnya yang mengkaji penaggulangan nilai hilang dilakukan oleh Nugraha, Prisyanto dan Pratama (2020), pada penelitian ini metode imputasi dilakukan pada data telemarketing adalah metode k-means. Hasil penelitian menunjukkan bahwa terjadi peningkatan akurasi setelah dilakukan penanggulangan terhadap permasalah missing value pada data tersebut.

Penelitian terdahulu yang dijabarkan sebelumnya memiliki konsep imputasi berbasis jarak. Imputasi ini memafaatkan jarak amatan yang akan di duga nilai nya kemudian diganti dengan nilai pemusatannya. Pada penelitian ini penulis mengusulkan penanggulangan *missing value* pada data BMKG dengan metode imputasi berbasis model (Model Based Imputation). Konsep imputasi ini adalah memanfaatkan nilai variabel lain untuk menduga nilai hilang, dengan mempertimbangkan hubungan antara variabel yang memiliki nilai hilang dengan variabel lain tersebut.

METODE PENELITIAN

Metode Regresi

Salah satu metode imputasi berbasis model untuk menduga nilai hilang pada variabel numerik adalah menggunakan metode regresi. Persamaan regresi yang ditunjukkan oleh formula berikut:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Maka model prediksinya dapat dituliskan sebagai berikut:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Nilai dugaan y ($\hat{y}$) digunakan untuk mengimputasi nilai hilang pada amatan. Pendugaan nilai parameter regresi dapat dilakukan dengan metode kuadrat terkecil.

Metode Regresi logistik

Variabel kategorik biner dapat diimputasi menggunakan regresi logistik. Analisis hubungan variabel respon yang berskala kategorik (2 kategori) dengan satu atau lebih variabel penjelas yang berskala kategorik atau kontinu dapat dilakukan dengan menggunakan regresi logistik biner(Hosmer dan Lemeshow, 2000). Kejadian variabel respon (Y) mengikuti sebaran Bernoulli dengan sebaran peluang sebagai berikut:

$$P(Y = y) = \pi^y (1 - \pi)^y$$

dengan y dapat bernilai 0 dan 1, π adalah peluang kejadian Y = 1. Kejadian peubah respon yang mengikuti sebaran Bernoulli yang dilakukan sebanyak n kali, dan setiap kejadian saling bebas maka peubah respon akan menyebar mengikuti sebaran Binomial. Secara umum, model regresi logistik dapat dijelaskan dengan persamaan berikut:

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}$$

Model regresi logistik biner diatas berbentuk non-linier, sehingga dapat diubah kedalam bentuk persamaan linear dengan cara melakukan transformasi logit. Berikut adalah bentuk persamaan yang telah ditransformasi dengan fungsi logit (Agresti, 1990):

$$logit\big(\pi(x)\big) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Untuk menduga parameter regresi pada persamaan diatas dapat dilakukan dengan metode kemungkinan maksimum (*Maximum Likelihood Estimator*) yakni menaksir parameter model regresi dengan cara memaksimumkan fungsi *likelihood* (Agresti, 2002). Berikut adalah bentuk fungsi *likelihood*:

$$L(\beta) = \prod_{i=1}^{n} \left[(\pi(x_i)^{y_i}) \big(1 - \pi(x_i)\big)^{1 - y_i} \right]$$

Random Forest (missForest)

MissForest merupakan teknik imputasi berbasis model non-parametrik dengan ide yang sama dengan pemodelan random forest. Metode ini dapat digunakan pada *mixed* variabel

yaitu variabel numerik maupun kategorik. Hasil pemodelan menggunakan random forest akan digunakan sebagai penduga amatan yang memiliki nilai hilang (Diouf dan Deme, 2022). Random Forest merupakan pengembangan metode Bagging (Bootstrap Aggregating). Metode ini bertujuan memperbaiki performa klasifikasi tunggal dengan nilai akurasi yang rendah (Wezel dan Potharst, 2007).

Algoritma imputasi dengan random forest (Hong dan Lynn, 2020) yaitu :

1) Inisialisasi. Untuk sebuah variabel yang mengandung missing values, maka akan digantikan dengan rata-rata data (untuk variabel kontinu) atau modus data (untuk data kategorik).
2) Imputasi. Variabel-variabel dari X_s dengan $s = 1, .., p$ yang mana diurutkan dari yang paling kecil sampai terbesar berdasarkan pada jumlah missing value-nya. Lalu missing value di imputasi untuk setiap X_s dan membangun sebuah model random forest menggunakan $y_{obs}^{(s)}$ sebagai respons dan $x_{obs}^{(s)}$ sebagai prediktor. Observasi-observasi yang ada dalam dataset dibagi menjadi dua bagian berdasarkan pada apakah data tersebut missing pada data sebenarnya. Jika memang bukan missing data pada data sebenarnya, maka data tersebut menjadi data training, sedangkan bila data tersebut missing pada data sebenarnya, maka data tersebut termasuk pada dataset yang diprediksi. Kemudian, missing value $y_{mis}^{(s)}$ bisa diprediksi menggunakan model random forest yang terbentuk pada $x_{mis}^{(s)}$. Hasil prediksi menggunakan model random forest mengantikan observasi yang bernilai missing pada data awal.
3) Berhenti. Saat semua missing data sudah terisi dengan hasil prediksi, maka iterasi Imputasi berhenti.

X_s adalah variabel ke-s yang mengandung missing value di $i_{miss}^{(s)} \subseteq \{1, \dots, n\}$. Kemudian $y_{obs}^{(s)}$ adalah nilai observasi dari X_s dan $y_{miss}^{(s)}$ adalah missing value dari X_s. Variabel-variabel lain X_s dengan $i_{obs}^{(s)} = \{1, \dots, n\}$. $i_{miss}^{(s)}$ dinotasikan dengan $x_{obs}^{(s)}$. Terlebih lagi variabel-variabel selain X_s yang memiliki observasi sesuai dengan $i_{mis}^{(s)}$ dinotasikan dengan $x_{miss}^{(s)}$.

Classification and Regression Tree (CART)

CART adalah teknik yang berkembang karena kemajuan teknologi komputer. Algoritma CART (Classification and Regression Trees) merupakan salah satu algoritma untuk menghasilkan pohon keputusan. Pada algoritma ini, sampel dibagi menjadi dua sub-sampel sehingga setiap simpul memiliki dua cabang. Pada CART sebagai model klasifikasi, pemisahan dilakukan menggunakan indeks Gini.

$$i(t) = 1 - \sum_{k=1}^{K} p^2(k|t)$$

Dimana $k|1, \ldots; K$ adalah indeks dari kelas, dan $p(k|t)$ adalah peluang bersyarat dari kelas k, asalkan kita berada di node t (Choubin et al., 2018).

Evaluasi Kebaikan model

Kebaikan model klasifikasi dapat dilihat dari *Confusion Matrix* (Tabel 1). *True Positive* adalah jumlah amatan yang tepat klasifikasi dari kelas positif, *True Negative* adalah jumlah amatan yang tepat klasifikasi dari kelas negatif, *False Negative* adalah jumlah amatan yang salah klasifikasi dari kelas positif sedangkan *False Positive* adalah jumlah amatan yang salah klasifikasi dari kelas negatif.

Tabel 1. *Confusion Matrix*

Prediksi	Aktual	
	Positif	Negatif
Positif	*True Positive*(TP)	*False Positive*(FP)
Negatif	*False Negative*(FN)	*True Negative*(TN)

Berdasarkan nilai pada *Confusion Matrix* dapat dihitung nilai *Accuracy, Sensitivity* dan *Specificity. Accuracy* menyatakan tingkat ketepatan *classifier* dalam mengklasifikan amatan. Berikut adalah formula yang digunakan dalam melihat performa klasifikasi :

$$Accuracy = \frac{TP + TN}{(TP + TN + FN + FP)}$$

Objek Penelitian dan Variabel Penelitian

Penelitian ini memanfaatkan data curah hujan harian di Provinsi Bengkulu dari tahun 01 Maret 2013 sampai dengan 27 Juli 2022. Data diperoleh dari seluruh stasiun Badan Meteorologi, Klimatologi dan Geofisika di Provinsi Bengkulu. Dengan rincian variabel yang digunakan adalah sebagai berikut:

Tabel 2. Penjabaran Skala Variabel

Jenis Variabel	Nama Variabel	Skala	Keterangan
Variabel Respon (Target)	Curah Hujan	Nominal	0 = Tidak Hujan 1 = Hujan
Variabel Prediktor	Suhu Rata-rata	Rasio	℃ = Derajat Celcius
	Suhu Maksimum	Rasio	℃ = Derajat Celcius

	Suhu Minimum	Rasio	℃ = Derajat Celcius
	Tekanan Udara diatas Permukaan Laut	Rasio	Hpa = Hecto Pascal
	Kecepatan Angin Maksimum	Rasio	Km/h
	Tingkat Awan	Nominal	"1/8", "2/8", "3/8", "4/8", "5/8", "6/8", "7/8", "8/8"
	Lama Penyinaran Matahari	Rasio	Hours (Jam)
	Arah Angin	Nominal	N = North (Utara) NE = NorthEast (Timur Laut) E = East (Timur) SE = SouthEast (Tenggara) S = South (Selatan) SW = SouthWest (Barat Daya) W = West (Barat) NW = NorthWest (Barat Laut)

Tahapan Analisis Data

Adapun langkah –langkah yang akan dilakukan dalam penelitian adalah sebagai berikut:

1. *Preprocessing data*, lakukan pengecekan data jika ada data yang kosong (*missing value*).
2. Penanggulangan missing value dengan cara mengganti nilai yang hilang dengan imputasi berbasis model
   a. Jika variabel numerik yang memuat nilai hilang maka digunakan metode regresi dan MissForest untuk mengganti nilai hilang
   b. Jika variabel kategorik yang memuat nilai hilang maka digunakan metode regresi logistik dan MissForest untuk mengganti nilai hilang
3. Melakukan pemodelan klasifikasi curah hujan dengan tahapan 4-6 pada data lengkap (setelah imputasi).
4. Membagi data menjadi data *training* dan *testing*, data *training* digunakan untuk pemodelan dan data *testing* digunakan untuk evaluasi performa klasifikasi. Data dibagi

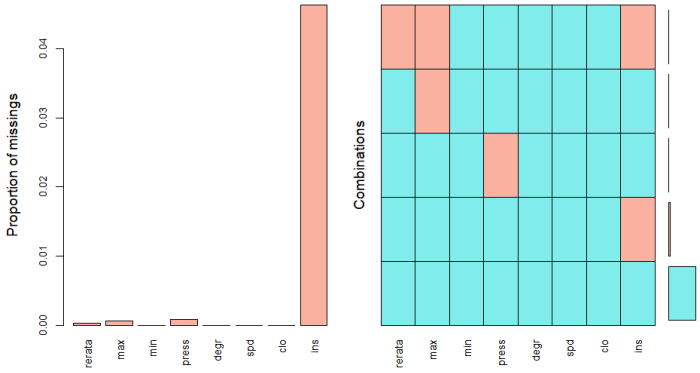
menjadi beberapa pilihan yakni dengan proporsi 70:30, 75:25 dan 80:20 dari total data yang tersedia.

5. Melakukan pemodelan klasifikasi *CART*
6. Melakukan evaluasi performa klasifikasi menggunakan nilai-nilai pada *confusion matrix*, data yang digunakan pada tahapan evaluasi adalah data testing

HASIL DAN PEMBAHASAN

Eksplorasi Data

Penelitian ini berfokus pada penanganan data hilang sehingga perlu diketahui pada variabel apa saja yang mengandung data hilang dan jumlah amatan dengan variabel yang tidak lengkap (hilang). Gambar 1 mengilustrasikan bahwa variabel penjelas lama penyinaran matahari memiliki nilai hilang paling banyak, kemudian variabel tekanan udara diatas permukaan laut, suhu rata-rata dan suhu maksimum.



Gambar 1. Plot Kondisi Data Hilang Dalam Dataset

Posisi nilai hilang setiap variabel pada dataset disajikan dalam Gambar 2. Berdasarkan gambar dapat dilihat bahwa terdapat 3268 amatan yang lengkap untuk semua variabel, 158 amatan memiliki nilai hilang pada lamanya penyinaran matahari, 1 amatan memiliki nilai hilang pada variabel tekanan udara diatas permukaan laut, 1 amatan memiliki nilai hilang pada variabel suhu maksimum, suhu rata-rata dan lamanya peyinaran matahari.

Penanganan Nilai Hilang

Pada penelitian, dilakukan penanganan nilai hilang dengan memprediksi nilai yang tidak lengkap atau hilang tersebut dengan memanfaatkan informasi dari variabel lainnya yang memiliki nilai lengkap. Metode yang digunakan adalah metode imputasi regresi dan imputasi random forest. Hasil akurasi model setelah permasalahan dataset ditanggulangi akan dinilai dengan melihat nilai akurasi. Pemodelan klasifikasi curah hujan dilakukan dengan metode CART dengan beberapa skenario pembagian dataset *training* dan *testing*

yaitu 70:30, 75:25 dan 80: 20.

Skenario pertama yaitu menggunakan 70% data untuk pemodelan dan 30% data untuk evaluasi model. Jika diperhatikan dari nilai *Accuracy, Sensitivity, Specificity* dan F1-*measure* pada Tabel 3 dapat disimpulkan bahwa tidak terlalu terdapat perbedaan akurasi model ketika imputasi dilakukan dengan random forest dan regresi. Demikian, juga jika dilihat dari kebaikan pemodelan antara data training dan testing sangat stabil sehingga dapat dikatakan model sudah baik namun belum maksimal.

Tabel 3. Ringkasan Kebaikan Model CART pada Skenario 70:30

Imputasi	70% (Training)				30% (Testing)			
	Accuracy	Sensitivity	Specificity	F-1	Accuracy	Sensitivity	Specificity	F-1
RF	0,6743	0,3293	0,9144	0,2267	0,6288	0,2723	0,9036	0,1949
Regresi	0,6782	0,3851	0,8905	0,2507	0,6441	0,3279	0,8666	0,2161

Skenario kedua yaitu menggunakan 75% data untuk pemodelan dan 25% data untuk evaluasi model. Sama halnya dengan skenario pertama, jika diperhatikan dari nilai Accuracy, Sensitivity, Specificity dan F1-measure pada Tabel 4 dapat disimpulkan bahwa tidak terlalu terdapat perbedaan akurasi model ketika imputasi dilakukan dengan random forest dan regresi. Demikian, juga jika dilihat dari kebaikan pemodelan antara data training dan testing sangat stabil sehingga dapat dikatakan model sudah baik namun belum maksimal.

Tabel 4 Ringkasan Kebaikan Model CART pada Skenario 75:25

Imputasi	75% (Training)				25% (Testing)			
	Accuracy	Sensitifity	Specificity	F-1	Accuracy	Sensitifity	Specificity	F-1
RF	0,6695	0,3841	0,8768	0,2472	0,6619	0,3959	0,8458	0,2446
Regresi	0,6705	0,3625	0,8904	0,2391	0,6469	0,3351	0,874	0,2222

Skenario ketiga yaitu menggunakan 75% data untuk pemodelan dan 25% data untuk evaluasi model. Sama halnya dengan skenario pertama dan kedua, jika diperhatikan dari nilai Accuracy, Sensitivity, Specificity dan F1-measure pada Tabel 5 dapat disimpulkan bahwa tidak terlalu terdapat perbedaan akurasi model ketika imputasi dilakukan dengan random forest dan regresi. Demikian, juga jika dilihat dari kebaikan pemodelan antara data training dan testing sangat stabil sehingga dapat dikatakan model sudah baik namun belum maksimal.

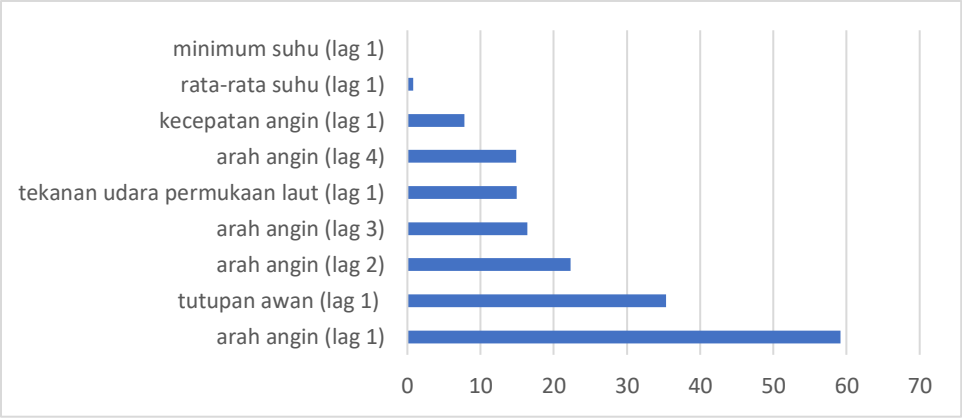
Tabel 5. Ringkasan Kebaikan Model CART pada Skenario 80:20

Imputasi	80% (Training)				20% (Testing)			
	Accuracy	Sensitifity	Specificity	F-1	Accuracy	Sensitifity	Specificity	F-1
RF	0,6786	0,4601	0,8331	0,2712	0,6434	0,4071	0,8247	0,2489
Regresi	0,668	0,4378	0,8322	0,2617	0,6566	0,4307	0,8232	0,2579

Berdasarkan perbandingan ketiga skenario pembagian data, dapat disimpulkan bahwa kedua metode imputasi yang digunakan dalam penelitian memberikan model yang stabil hal ini dapat dilihat dari kebaikan model yang hampir sama antara data training dan data testing. Nilai aku dapat dikatakan belum cukup tinggi hal ini dapat dikarenakan oleh variabel prediktor yang digunakan dalam model belum sepenuhnya dapat memisahkan amatan kategori hujan dan tidak hujan. Pada peelitian selanjutnya akan dipertimbangkan menambah variabel lain ke dalam model.

Kontribusi Variabel Prediktor terhadap Model

Kontribusi variabel prediktor terhadap pemodelan klasifikasi dengan menggunakan CART dapat dilihat dari Gambar 2. Variabel yang memiliki kontribusi paling besar adalah arah angin pada hari sebelumnya, kemudian diikuti berturut-turut oleh variabel tutupan awan pada satu hari sebelumnya, arah angin dua hari sebelumnya. Variabel dengan kontribusi paling kecil adalah suhu, baik rata-rata suhu pada satu hari sebelumnya maupun minimum suhu pada satu hari sebelumnya.



Gambar 2. Kontribusi Setiap Variabel Dalam Model.

SIMPULAN

Berdasarkan uraian pada subbab sebelumnya dapat disimpulkan bahwa:

1. Kedua metode imputasi berbasis model yang digunakan memberikan nilai akurasi yang tidak terlalu berbeda.

2. Nilai akurasi konsiten antara data training dan data testing, hal ini mengindikasikan bahwa imputasi yang dilakukan sudah baik.
3. Pemodelan dilakukan dengan menggunakan metode CART dengan variabel dengan kontribusi tinggi adalah variabel yang memiliki kontribusi paling besar adalah arah angin pada hari sebelumnya dan variabel tutupan awan pada satu hari sebelumnya.

DAFTAR PUSTAKA

Acuna, E., dan Rodriguez, C. (2004). The Treatment of Missing Values and Its Effect on Classifier Accuracy, Classification, Clustering, and Data Mining Applications. Springer Berlin Heidelberg, hal. 639-647.

Agresti A. (1990). Categorical Data Analysis. New Jersey : John Wiley and Sons. 558

Agresti, A. (2002). Categorical Data Analysis, John Wiley and Sons, Inc Second Edition. New York. Alam, M.S.,

Fadillah, I. J., & Puspita, C. D. (2020). Pemanfaatan metode weighted k-nearest neighbor imputation (weighted knni) untuk mengatasi missing data. In Seminar Nasional Official Statistics (Vol. 2020, No. 1, pp. 511-518).

Hosmer D.W., Lemeshow S. (2000). Applied Logistic Regression, 2nd edition. New York : John Wiley and Sons. 373 23

James, G. Daniella, W. Trevor, H. and Robert, T. (2013). An Introduction of Statistical Learning. Springer :New York.

Kowarik, A. dan Templ, M. (2016) "Imputation with the R package VIM," Journal of Statistical Software, 74(7). doi:10.18637/jss.v074.i07.

Laencina, G., Go'mez, S., Vidal, F., dan Verleysen, M. (2009). K Nearest Neighbours with Mutual Information for Simultaneous Classification and Missing Data Imputation. Neurocomputing, Vol.72, hal. 1483–1493.

Nugraha, A. F., Pristyanto, Y., & Pratama, I. (2020). Penanganan Missing Values Untuk Meningkatkan Kinerja Model Machine Learning Pada Data Telemarketing. Pseudocode, 7(2), 165-171.

Wezel M.V., Potharst R. (2007). Improved Customer Choice Predictions using Ensemble Methods. European Journal of Operational Research, 181, 436-452. 24

Diouf, S., & Dème, A. (2022). Imputation methods for missing values: the case of Senegalese meteorological data. African Journal of Applied Statistics, 9(1), 1245-1278.