Xplore (2023), Vol. 12, No. 1, 22-35
DOI: https://doi.org/ 10.29244/xplore.v12i1.1056
Xplore: Journal of Statistics
Available at https://journal.stats.id/index.php/xplore
e-ISSN: 2655-2744

Perbandingan Metode Hot-deck, Regression dan K-Nearest Neighbor Imputation dalam Pendugaan Data Hilang pada Dapodik Tahun 2020*

Inayatul Izzati Diana Yusuf[1], Budi Susetyo[1‡] , La Ode Abdul Rahman[1]

[1]Department of Statistics, IPB University, Indonesia
‡corresponding author: budisu@apps.ipb.ac.id

Abstract

Data Pokok Pendidikan (Dapodik) is a nation-wide data collection system that contains data on education units. Missing value in Dapodik cause the loss of important information.to solve this problem can use imputation. Imputation is a procedure to predict the missing value with a certain method. This study aims to compare three imputation methods which are Hot-deck imputation, Regression Imputation and K-Nearest Neighbor imputation (KNNI). Hot-deck imputation is an imputation method with values that have similar characteristics. Regression imputation is a method to predict missing value by using regression approach. KNNI is an imputation method which groups data based on the closest neighbor. Simulation for generating missing value was carried out by dividing the percentage of 2%, 3%, 4% and 5%, then imputed with the three methods. The best model is determined based on the lowest value of Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The best imputation method based on the lowest RMSE and MAPE values is a regression imputation.

Keywords: dapodik, hot-deck imputation, KNNI, missing value, regression imputation

1. Pendahuluan

Pemanfaatan sistem informasi pada data pendidikan dilakukan guna merealisasikan perencanaan program pendidikan yang tepat sasaran. Oleh karena itu, Kementerian Pendidikan Kebudayaan Riset dan Teknologi (Kemendikbudristek) mengembangkan sistem pendataan skala nasional yang disebut dengan Data Pokok Pendidikan (Dapodik). Dapodik adalah suatu sistem pendataan yang memuat data satuan pendidikan, pendidik, tenaga kependidikan, peserta didik, dan substansi pendidikan yang bersumber dari satuan pendidikan dan terus menerus diperbarui secara online

* Received: Okt 2022; Reviewed: Des 2022; Published: Jan 2023

(Permendikbud Nomor 79 Tahun 2015). Dapodik memerlukan data yang berkualitas untuk proses pelaksanaan dan penilaian program kerja pendidikan nasional.

Menurut Mosley (2008) dalam bukunya “Dictionary of Data Management” Kualitas data adalah level data yang menyatakan data tersebut akurat (accurate), lengkap (complete), update, dan konsisten. Salah satu masalah yang relevan dalam kualitas data adalah adanya data hilang atau missing value. Suatu dataset yang didalamnya terdapat satu atau lebih atribut tidak memiliki data disebut dataset yang memiliki data hilang (Garcia et al. 2010). Little dan Rubin (1987) menjelaskan terdapat tiga tipe data hilang, yaitu Missing Completely At Random (MCAR), Missing At Random (MAR) dan Not Missing At Random (NMAR). Menurut Enders (2010) MCAR merupakan tipe data hilang yang benar-benar hilang secara acak, terjadi ketika probabilitas hilangnya data pada suatu peubah tidak terkait dengan peubah lain dan juga tidak terkait dengan peubah itu sendiri. Selanjutnya MAR merupakan tipe data hilang ketika probabilitas hilangnya data pada suatu peubah terkait dengan satu atau lebih peubah lainnya, tetapi tidak dengan nilai pada peubah itu sendiri. Sedangkan NMAR merupakan tipe data hilang ketika probabilitas hilangnya data pada suatu peubah terkait dengan nilai itu sendiri.

Little dan Rubin (2002) memperkenalkan beberapa prosedur untuk mengatasi data hilang yang salah satunya adalah imputasi. Imputasi merupakan prosedur untuk memprediksi nilai yang hilang berdasarkan peubah lainnya dengan metode tertentu kemudian mengisi (fill in) data yang hilang dengan nilai tersebut. Jerez dan Molina (2010) menjelaskan bahwa metode imputasi dikelompokkan menjadi dua jenis, yaitu metode imputasi berbasis statistik dan machine learning. Pada penelitian ini akan membandingkan penerapan metode imputasi hot-deck imputation dan regression imputation yang termasuk contoh metode imputasi berbasis statistik, dan menggunakan salah satu metode imputasi berbasis machine learning yaitu K-Nearest Neighbor Imputation (KNNI).

Fadillah dan Muchlisoh (2017) telah melakukan penelitian menggunakan dua dataset lengkap awal yaitu data hasil bangkitan dan data Susenas Kor dan Konsumsi Maret 2017. Setelah itu dilakukannya simulasi pembangkitan data hilang dengan tipe data hilang MCAR. Dataset yang mengandung data hilang tersebut diimputasi menggunakan metode hot-deck imputation dan KNNI. Penelitian tersebut menghasilkan kesimpulan bahwa metode KNNI mampu menduga data hilang lebih baik. Penelitian pendugaan data hilang juga dilakukan oleh Rhaudatunnisa dan Wilantika (2021) menggunakan lima dataset hasil simulasi pembangkitan data hilang dengan tipe data MCAR, MAR dan NMAR berdasarkan lima dataset lengkap awal, yaitu Susenas Kor Maret 2019 dataset, Iris dataset, E.Coli dataset, Breast Cancer 1 dataset, dan Breast Cancer 2 dataset. Kemudian dilakukan imputasi menggunakan perbandingan tiga metode imputasi, yaitu hot-deck imputation, k-nearest neighbor imputation, dan Predictive Mean Matching (PMM) dengan analisis skoring. Penelitian tersebut menghasilkan kesimpulan bahwa metode hot-deck imputation memberikan skor tertinggi dan menjadi metode terbaik dalam menangani data hilang.

2. Metodologi

2.1 Data

Data yang digunakan dalam penelitian adalah data sekunder jenjang SMP di provinsi Jawa Barat yang bersumber dari laman Dapodik tahun 2020 kemudian diolah menjadi peubah-peubah yang tertera pada tabel 1. Data terdiri atas 14 peubah dengan jumlah 1422 sekolah. Peubah mencakup data guru, tenaga kependidikan, peserta didik dan rombongan belajar.

Tabel 1 Peubah yang digunakan

Peubah	Keterangan	Skala
Sekolah	Nama sekolah	Nominal
Kab/kota	Lokasi sekolah	Nominal
$\boldsymbol{X_1}$	Persentase lulusan	Rasio
$\boldsymbol{X_2}$	Persentase siswa *Drop Out* (DO)	Rasio
$\boldsymbol{X_3}$	Rasio jumlah siswa per rombongan belajar	Rasio
$\boldsymbol{X_4}$	Rasio jumlah guru per jumlah siswa	Rasio
$\boldsymbol{X_5}$	Persentase guru memiliki sertifikat	Rasio
$\boldsymbol{X_6}$	Persentase guru sudah S1	Rasio
$\boldsymbol{X_7}$	Rasio tenaga administrasi per jumlah rombongan belajar	Rasio
$\boldsymbol{X_8}$	Rasio jumlah ruang kelas per jumlah rombongan belajar	Rasio
$\boldsymbol{X_9}$	Rasio jumlah komputer per jumlah siswa	Rasio
$\boldsymbol{X_{10}}$	Rasio jumlah siswa per jumlah WC	Rasio
$\boldsymbol{X_{11}}$	Rasio ketersediaan lab IPA, komputer dan bahasa	Rasio
$\boldsymbol{X_{12}}$	Rasio ketersediaan ruang penunjang terdiri atas ruang osis, ruang perpustakaan dan fasilitas olahraga	Rasio

2.2 Metode Penelitian

Proses analisis pada penelitian ini menggunakan software R Studio *(package: missMethods, StatMatch, tibble, caret, missForest, dplyr, impute, VIM*) dan MS.Excel. Tahapan - tahapan yang dilakukan dalam penelitian ini sebagai berikut:

1. Melakukan simulasi dengan membangkitkan data hilang sebesar 2%, 3%, 4% dan 5% pada masing-masing tipe data *Missing Completely At Random* (MCAR), *Missing At Random* (MAR) dan *Not Missing At Random* (NMAR):
   a. MCAR yaitu membangkitkan data hilang pada setiap peubah $X_1 - X_{12}$ secara acak tanpa terkait peubah lain dan peubah itu sendiri
   b. MAR yaitu membangkitkan data hilang pada setiap peubah $X_1 - X_{12}$ dengan ketentuan terkait dengan peubah lain yaitu data hilang dibangkitkan pada sekolah yang berada di wilayah kabupaten
   c. NMAR yaitu membangkitkan data hilang pada setiap peubah $X_1 - X_{12}$ dengan ketentuan terkait dengan peubah itu sendiri yaitu data hilang dibangkitkan pada observasi yang memiliki nilai dibawah rata-rata setiap peubahnya

2. Melakukan pendugaan data hilang menggunakan metode *hot-deck imputation.* Langkah-langkah yang dilakukan dalam *hot-deck imputation* meliputi:
   a. Mengelompokkan observasi data hilang kedalam masing-masing kelompok yang dianggap memiliki karakteristik serupa, yaitu berdasarkan wilayah sekolah
   b. Menghitung jarak *euclidean* pada masing-masing kelompok tersebut sesuai dengan persamaan berikut (Wilson dan Martinez 1997):

$$d_{ij} = \sqrt{\sum_{\mathrm{p}=1}^{n}\left(x_{ip} - x_{jp}\right)^2} \quad (1)$$

   Keterangan:

   d_{ij} = jarak *Euclidean* observasi ke-*i* terhadap observasi ke-*j*
   n = banyak peubah
   x_{ip} = data observasi ke-i pada peubah ke-*p*
   x_{jp} = data observasi ke-j pada peubah ke-*p*

   c. Melakukan proses imputasi data hilang menggunakan nilai observasi yang memiliki jarak *euclidean* terkecil
3. Melakukan pendugaan data hilang menggunakan metode *regression imputation.* Langkah-langkah yang dilakukan dalam *regression imputation* meliputi:
   a. Mengidentifikasi peubah bebas (X) dan peubah terikat (Y). Peubah terikat (Y) yaitu peubah yang mengandung data hilang
   b. Menduga model regresi linear berganda berdasarkan data lengkap sesuai dengan persamaan berikut (Mostafa 2019):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_i X_i \quad (2)$$

   Keterangan:

   Y = peubah terikat
   X_i = peubah bebas
   β_0 = konstanta
   β_1 = koefisien regresi peubah ke-*i*

   c. Melakukan imputasi data hilang menggunakan persamaan regresi yang telah dicari sebelumnya pada setiap peubahnya
4. Melakukan pendugaan data hilang menggunakan metode KNNI. Langkah yang dilakukan dalam metode KNNI meliputi :
   a. Menentukan parameter *k* dengan persamaan $k = \sqrt{n}$
   b. Menghitung jarak *euclidian* sesuai dengan persamaan berikut (Wilson dan Martinez 1997):

$$d_{ij} = \sqrt{\sum_{\mathrm{p}=1}^{n}\left(x_{ip} - x_{jp}\right)^2} \quad (3)$$

Keterangan :

d_{ij} = jarak *Euclidean* observasi ke-i terhadap observasi ke-*j*

n = banyak peubah

x_{ip} = data observasi ke-i pada peubah ke-*p*

x_{jp} = data observasi ke-j pada peubah ke-*p*

c. Mengurutkan hasil perhitungan jarak *euclidean* dari yang terkecil dan mengambil sebanyak jumlah *k* yang telah ditentukan

d. Melakukan proses imputasi data hilang sesuai jumlah *k* yang telah ditentukan dengan menghitung nilai *weight mean estimation* (WME) berdasarkan persamaan berikut (Sallaby dan Azlan 2021) :

$$\overline{x_j} = \frac{\sum_{k=1}^{K} w_k v_k}{\sum_{k=1}^{K} w_k} \quad (4)$$

Keterangan :

$\overline{x_j}$ = estimasi rata-rata

v_k = nilai data lengkap pada peubah data hilang

w_k = bobot observasi tetangga terdekat ke-k = $\frac{1}{(d_{ij})^2}$

K = jumlah observasi terdekat yang digunakan

k = observasi dari K

5. Menghitung nilai *Root Mean Square Error* (RMSE) sebagai tolak ukur *error rate* pada setiap metode imputasi sesuai dengan persamaan berikut (Jerez dan Molina 2010) :

$$\text{RMSE} = \sqrt{\frac{1}{\text{n}}\sum_{\text{i}=1}^{n}(\hat{\text{y}}_i - y_i)^2} \quad (5)$$

Keterangan :

$\hat{y}_i$ = nilai hasil imputasi observasi ke-i

y_i = nilai aktual observasi ke-i

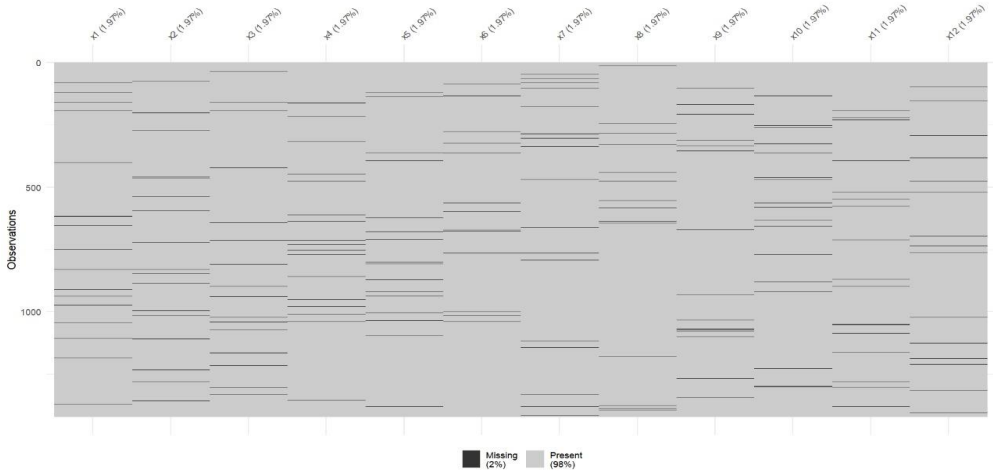
n = banyaknya data hilang

6. Menghitung nilai *Mean Absolute Percentage Error* (MAPE) sebagai tolak ukur *error rate* pada setiap metode imputasi sesuai dengan persamaan berikut (Montgomery *et al.* 2015) :

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i}\ x\ 100\% \quad (6)$$

Keterangan:

$\hat{y}_i$ = nilai hasil imputasi observasi ke-i



y_i = nilai aktual observasi ke-i

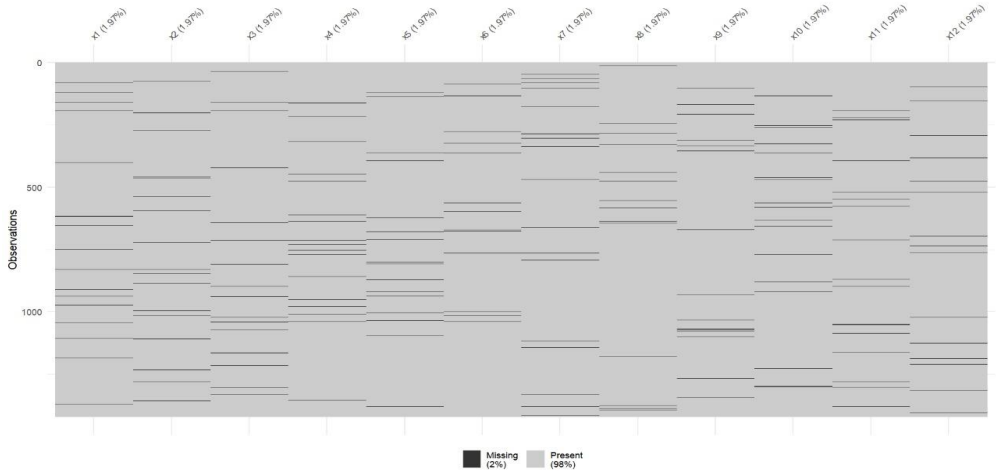
n = banyaknya data hilang

Menentukan metode imputasi yang lebih baik digunakan dengan mengevaluasi hasil perhitungan nilai RMSE dan MAPE.

3. Hasil dan Pembahasan

3.1 Simulasi Data Hilang

Simulasi dilakukan dengan membangkitkan data hilang sebesar 2% yaitu 28 observasi, 3% yaitu 43 observasi, 4% yaitu 57 observasi dan 5% yaitu 71 observasi pada masing-masing tipe data hilang yaitu MCAR, MAR dan NMAR. Sebaran posisi data hilang hasil simulasi akan berbeda-beda untuk setiap tipe data dan persentase yang dibangkitkan. Berikut adalah salah satu contoh sebaran data hilang pada tipe data MCAR 2% yang dapat dilihat pada Gambar 1.



Gambar 1 Sebaran posisi data hilang MCAR2%

3.2 Pendugaan Data Hilang Metode *Hot-deck Imputation*

Hot-deck imputation adalah metode imputasi untuk mengatasi data hilang dengan nilai yang memiliki karakteristik serupa dari observasi lain dan diterapkan pada nilai data hilang (Scheaffer *et al.* 2012). Karakteristik serupa didapatkan dengan mengelompokkan observasi data hilang kedalam masing-masing kelompok yang memiliki wilayah sama. Kemudian menentukan tetangga terdekat menggunakan perhitungan jarak (Farhangfar *et al.* 2008). Ukuran jarak yang digunakan adalah *euclidean distance* dan menggunakan observasi yang menghasilkan nilai *euclidean* terkecil untuk melakukan imputasi. Berikut adalah tabel hasil perhitungan jarak euclidean pada salah satu peubah, yaitu peubah persentase guru memiliki sertifikat (X_5) observasi hilang 121 data MCAR 2%:

Tabel 2 Jarak *euclidean* peubah $\mathbf{X_5}$ observasi 121

No.	Wilayah	Observasi	Nilai imputasi	Jarak *euclidean*
1	Kab. Ciamis	OBS_{121}, OBS_{31}	84,62	0,879
2	Kab. Ciamis	OBS_{121}, OBS_{36}	0,00	2,814
3	Kab. Ciamis	OBS_{121}, OBS_{56}	71,43	43,416
4	Kab. Ciamis	OBS_{121}, OBS_{63}	75,00	31,517
5	Kab. Ciamis	OBS_{121}, OBS_{87}	57,89	29,87
…	…	…	…	…
…	…	…	…	…
70	Kab. Ciamis	OBS_{121}, OBS_{1409}	62,5	33,897

Observasi ke-121 merupakan sekolah SMP Negeri Satu Atap Cipaku yang berada di Kabupaten Ciamis. Oleh karena itu dilakukan pengelompokkan seluruh sekolah yang berada di Kabupaten Ciamis yang hasilnya terdapat 70 sekolah. Setelah itu dilakukan penghitungan jarak *euclidean* pada masing-masing observasi lengkap lain pada kelompok tersebut.Berdasarkan tabel, didapatkan hasil bahwa nilai *euclidean* terkecil dihasilkan oleh observasi ke-31 yaitu 0,879 dan mendapatkan nilai imputasi yaitu 84,62 untuk mengisi data hilang pada observasi 121 peubah persentase guru memiliki sertifikat (X_5). Berikut merupakan tabel perbandingan antara data asli dan data hasil imputasi dengan mengambil contoh lima observasi hilang pada salah satu tipe data yaitu MCAR2% peubah persentase guru memiliki sertifikat (X_5) menggunakan *hot-deck imputation*:

Tabel 3 Perbandingan data MCAR 2% peubah $\boldsymbol{X_5}$ metode *hot-deck*

Obs	Data asli $(\boldsymbol{y_i})$	Hasil dugaan $(\boldsymbol{\hat{y}_i})$	Selisih $\lvert \boldsymbol{y_i} - \boldsymbol{\hat{y}_i} \rvert$
121	54,55	84,62	30,07
138	76,09	44,00	32,09
323	51,35	30,30	21,05
362	73,08	60,00	13,08
389	90,32	29,17	61,15

3.3 Pendugaan Data Hilang Metode *Regression Imputation*

Metode *regression imputation* adalah metode untuk memprediksi data hilang menggunakan pendekatan regresi. Regresi linear berganda berfungsi untuk melihat pengaruh hubungan antara satu atau lebih peubah bebas (X) dengan satu peubah terikat (Y) yang bersifat linear (Hastie *et al*. 2008). Pada penelitian ini, setiap peubah di setiap dataset akan dicari fungsi persamaan linearnya terlebih dahulu untuk masing-masing peubahnya, selanjutnya hasil dugaan persamaan regresi digunakan untuk mengisi nilai yang hilang. Berikut merupakan salah satu persamaan linear yang didapatkan pada peubah persentase guru memiliki sertifikat (X_5) *dataset* MCAR 2%:

$\widehat{X_5}$= - 38.96554 + 0.23543X_1 + 0.38539X_2+ 0.42375X_3 - 24.18866X_4 + 0.48140X_6 - 3.76130X_7 - 5.36891X_8 + 51.68884X_9 + 0.02154X_{10} + 23.62106X_{11} + 20.26855X_{12}

Selanjutnya menggunakan persamaan linear tersebut dilakukannya imputasi untuk mengisi nilai yang hilang. Perbandingan data asli dengan data hasil imputasi dengan mengambil contoh lima observasi hilang pada salah satu tipe data yaitu MCAR2% peubah persentase guru memiliki sertifikat (X_5) menggunakan *regression imputation* dapat dilihat pada tabel berikut:

Tabel 4 Perbandingan data MCAR 2% peubah $\boldsymbol{X_5}$ metode *regression*

Obs	Data asli (y_i)	Hasil dugaan $(\hat{y}_i)$	Selisih $\lvert y_i - \hat{y}_i \rvert$
121	54,55	28,21	26,34
138	76,09	52,13	23,96
323	51,35	65,95	14,60
362	73,08	39,61	33,47
389	90,32	54,08	36,24

3.4 Pendugaan Data Hilang Metode *K-Nearest Neighbor Imputation* (KNNI)

Metode *K-Nearest Neighbor Imputation* (KNNI) mengklasifikasikan data berdasarkan keanggotaan tetangga terdekat (Avelita 2013). Sehingga perlu menentukan jumlah tetangga terdekat (*k)* untuk selanjutnya dilakukan imputasi. Pada penelitian ini nilai *k* yang digunakan adalah *k* = 37 berdasarkan persamaan *k* = √n (Novita et al. 2018) dengan jumlah data hilang 2%, 3%, 4% dan 5% untuk setiap tipe data. Setelah itu dihitung jarak *euclidean* antara observasi yang mengandung data hilang dengan observasi lengkap, lalu diurutkan nilainya dari yang terkecil hingga terbesar dan mengambil sebanyak jumlah *k* yang telah ditentukan.

Pada data MCAR 2% peubah persentase guru memiliki sertifikat (X_5) terdapat 28 data yang dihilangkan yaitu observasi ke 121, 138, 323 hingga 1418 yang untuk lengkapnya bisa dilihat pada tabel 10. Berikut adalah 37 jarak tetangga terdekat menggunakan jarak euclidean antara observasi ke-121 dengan observasi lengkap:

Tabel 5 Jarak *euclidean* peubah $\boldsymbol{X_5}$ observasi 121 berdasarkan *rank*

Observasi	*Ranking*	Jarak e*uclidean*
OBS_{121}, OBS_{31}	1	0,879
OBS_{121}, OBS_{1406}	2	0,892
OBS_{121}, OBS_{819}	3	0,960
OBS_{121}, OBS_{569}	4	0,965
OBS_{121}, OBS_{18}	5	0,970
OBS_{121}, OBS_{654}	6	1,089
OBS_{121}, OBS_{776}	7	1,100
…	...	...
…	…	…
OBS_{121}, OBS_{831}	37	2,175

Tabel 9 menunjukkan nilai hasil perhitungan jarak *euclidean* yang sudah diurutkan mulai dari yang terkecil hingga terbesar untuk peubah persentase guru memiliki sertifikat (X_5) observasi ke-121 dataset MCAR 2%. Data hilang terdapat pada observasi ke-121 dengan jarak *euclidean* terkecil yaitu pada observasi ke-31 dengan nilai 0,879 lalu jarak terkecil kedua pada observasi ke-1406 dengan nilai 0,892 hingga jarak *euclidean* urutan ke-37 pada observasi ke-831 dengan nilai 2,175. Selanjutnya dilakukan imputasi berdasarkan jumlah *k=37* yang sudah ditentukan dengan *weight mean* estimation sesuai dengan persamaan 4.

$$\bar{X}_{121} = \frac{\frac{1}{(0{,}879)^2}84{,}62}{\frac{1}{(0{,}879)^2}} + \frac{\frac{1}{(0{,}892)^2}25}{\frac{1}{(0{,}892)^2}} + \frac{\frac{1}{(0{,}96)^2}25}{\frac{1}{(0{,}96)^2}} + \ldots + \frac{\frac{1}{(2{,}175)^2}14{,}28}{\frac{1}{(2{,}175)^2}} = 36{,}38$$

Hasil imputasi pada observasi ke-121 menggunakan *k* = 37 yaitu 36,38. Hasil imputasi data hilang pada observasi lain dapat dilakukan dengan cara yang sama seperti observasi ke-121 peubah X_5. Berikut disajikan perbandingan antara data asli dan data hasil imputasi dengan mengambil contoh lima observasi pada salah satu tipe data yaitu MCAR2% peubah persentase guru memiliki sertifikat (X_5) menggunakan metode KNNI:

Tabel 6 Perbandingan data MCAR 2% peubah $\boldsymbol{X_5}$ metode KNNI

Obs	Data asli (y_i)	Hasil dugaan $(\hat{y}_i)$	Selisih $\lvert y_i - \hat{y}_i \rvert$
121	54,55	36,38	18,17
138	76,09	63,61	12,48
323	51,35	64,96	13,61
362	73,08	48,70	24,38
389	90,32	61,63	28,69

3.5 Akurasi dan Evaluasi Model

Akurasi dan evaluasi model dilakukan sebagai tolak ukur untuk menentukan metode yang lebih baik digunakan. Model yang telah didapatkan akan diuji menggunakan parameter RMSE dan MAPE. RMSE dipilih karena merupakan yang paling umum dan memiliki sensitivitas yang cukup tinggi sehingga akurasinya baik, Sedangkan MAPE, dipilih karena memberikan persentase kesalahan hasil sehingga lebih akurat dan mudah di interpretasikan

Tabel 7 Nilai akurasi hasil imputasi tipe data MCAR

Metode	RMSE				MAPE(%)			
	2%	3%	4%	5%	2%	3%	4%	5%
Hot-deck	17,21	20,72	13,88	19,30	56,08	56,34	37,29	55,06
Regression	8,47	9,69	9,10	9,42	37,67	41,05	44,01	45,03
KNNI	9,02	9,88	9,27	9,61	37,02	39,84	47,40	52,75

Tabel 13 menunjukkan nilai akurasi pada data MCAR. Pada tabel terlihat bahwa untuk tipe data MCAR, secara umum metode *regression imputation* mampu menduga nilai data hilang lebih baik dibandingkan kedua metode lainnya. Hal tersebut terlihat dari nilai RMSE yang dihasilkan paling rendah untuk semua persentase data hilang. Sedangkan untuk MAPE dihasilkan nilai yang cukup variatif. Metode KNNI menghasilkan nilai terendah pada persentase 2% yaitu 37,02 dan pada persentase 3% yaitu 39,84. Pada persentase 4% dihasilkan nilai terendah oleh *hot-deck imputation* yaitu 37,29 dan pada persentase 5% dihasilkan nilai terendah oleh *regression imputation* yaitu 45,03*.* Metode *regression imputation* mampu melakukan imputasi data hilang dengan cukup stabil antar persentasenya*.* Sedangkan untuk metode *hot-deck imputation* menghasilkan nilai yang cukup jauh berbeda sehingga dikatakan kurang baik dibandingkan kedua metode lain pada tipe data MCAR, *hotdeck-imputation* juga sangat bergantung posisi data hilang sehingga nilainya dapat berubah signifikan contohnya pada persentase data hilang 4%.

Tabel 8 Nilai akurasi hasil imputasi tipe data MAR

Metode	RMSE				MAPE(%)			
	2%	3%	4%	5%	2%	3%	4%	5%
Hot-deck	19,29	25,94	15,47	16,85	48,84	58,00	63,76	53,24
Regression	11,92	7,94	10,21	12,93	44,16	44,90	49,58	52,91
KNNI	11,29	7,83	9,84	12,61	42,39	47,53	51,41	56,37

Tabel 14 menunjukkan nilai akurasi hasil imputasi untuk tipe data MAR. Nilai RMSE pada metode KNNI menghasilkan nilai terendah pada semua persentase data hilang. Sedangkan untuk perbandingan MAPE, Pada persentase 2% diperoleh nilai MAPE terendah yaitu 42,39% dihasilkan oleh metode KNNI dan mengalami perubahan pada persentase 3%, 4% dan 5% dihasilkan nilai MAPE terendah oleh metode *regression.* Dapat disimpulkan secara keseluruhan metode yang lebih baik dalam

menduga data hilang pada tipe data MAR yaitu metode KNNI. Pada tipe data ini, terlihat bahwa pada persentase 5% semua metode menghasilkan nilai MAPE > 50% sehingga model sudah tidak dapat digunakan lagi.

Tabel 9 Nilai akurasi hasil imputasi tipe data NMAR

Metode	**RMSE**				**MAPE(%)**			
	2%	3%	4%	5%	2%	3%	4%	5%
Hot-deck	16,01	25,41	27,07	11,42	41,16	35,08	44,47	61,58
Regression	10,35	21,42	21,81	9,26	33,54	32,58	39,29	59,41
KNNI	9,59	20,18	21,00	10,23	34,81	36,91	41,16	67,73

Tabel 15 menunjukkan akurasi untuk tipe data NMAR. Jika dilihat berdasarkan nilai RMSE metode KNNI secara umum memiliki nilai RMSE terendah waluapun mengalami perubahan pada persentase data hilang 5%, namun nilainya hanya memiliki selisih sedikit saja. Sedangkan jika dilihat dari nilai MAPE yang didapatkan, metode *regression imputation* menghasilkan nilai yang terendah pada semua persentase data hilang. Sehingga secara umum dapat dikatakan pada tipe NMAR metode *regression imputation* menjadi metode terbaik dibandingkan metode lainnya dalam melakukan imputasi.

4. Simpulan dan Saran

Berdasarkan pembahasan yang telah dilakukan, dapat disimpulkan bahwa, pendugaan data hilang pada Dapodik jenjang SMP di Jawa Barat tahun 2020 menunjukkan bahwa secara keseluruhan dari sisi akurasi metode *regression imputation* menjadi metode yang lebih baik digunakan diantara kedua metode lainnya pada hampir semua tipe data hilang dengan metode KNNI yang mempunyai selisih sedikit nilai pada saat dilakukan perbandingan evaluasi hasil imputasi, sehingga dapat dikatakan metode KNNI mampu menduga nilai data hilang sama baiknya dengan metode *regression imputation*. Sedangkan untuk metode *hot-deck imputation* disimpulkan kurang baik dalam menduga nilai data hilang pada Dapodik SMP di Jawa Barat tahun 2020.

Faktor yang mempengaruhi nilai RMSE dan MAPE salah satunya yaitu posisi observasi yang mengalami data hilang hasil simulasi. Pada penelitian ini juga terlihat bahwa nilai RMSE dan MAPE yang didapatkan dipengaruhi besarnya persentase data hilang yang digunakan, secara umum semakin meningkatnya jumlah data yang hilang, akan semakin besar pula nilai akurasi, yang diartikan pendugaan data hilang kurang

baik. Metode imputasi yang digunakan pada penelitian ini hanya mampu menduga nilai data hilang hingga persentase 4%.

Daftar Pustaka

Avelita B.2013. Klasifikasi_K-Nearest_Neighbor. [Dipetik 2016 06 22]. https://www.academia.edu/9131959/A._Klasifikasi_K- Nearest_Neighbor .

Enders CK.2010. *Applied Missing Data Analysis.* New York (US) : The Guilford Press

Fadillah IJ, Muchlisoh S.2017. Perbandingan Metode Hot-Deck Imputation Dan Metode Knni Dalam Mengatasi Missing Values. *Seminar Nas. Off. Stat*.2019(1):275–285.DOI:10.34123/semnasoffstat.v2019i1.101.

Farhangfar A, Kurgan L, Dy J.2008. Impact of imputation of missing values on classification for discreate data. *Pattern Recognition* 41(12):3692-3705.

Garcia LPJ, Sancho GJL, Figueiras VAR. 2010. Pattern classification with missing data : a review. *Neural Comput & Applic*. 19:263–282.

Hastie T, Tibshirani R, Friedman J.2008. *The elements of statistical learning: data mining, inference, and prediction 2 nd edition*. California: Springer..

Jerez JM, Molina I. 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*: 105-115.

Little RJA, Rubin DB.1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Son Inc.

Little, RJA, Rubin DB. 2002. *Statistical analysis with missing data, 2 nd edition*: Wiley:US.

Montgomery DC, Jennings CL, Kulahci M.2015. Introduction to Time Series Analysis an Forecasting. Canada.

Mosley M. 2008. *The DAMA Dictionary of Data Management 1st Edition*. USA:Technics Publications, LLC.

Mostafa SM.2019. Imputing missing values using cumulative linear regression. *Journals The Institution of Engineering and Technology*(4):182-200. DOI:10.1049.trit.2019.0032.

Novita S, Harsani P, Qur'ania R.2018. Penerapan K-Nearest Neighbor (KNN) Untuk Klasifikasi Anggrek Berdasarkan Karakter Morfologi Daun dan Bunga. 15(1):118-125.

[Permen] Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 79 Tahun 2015 Tentang Data Pokok Pendidikan. 2015.

Rhaudatunnisa T, Wilantika N. 2021. Performance Comparison of Hot-Deck Imputation, K-Nearest Neighbor Imputation, and Predictive Mean Matching in Missing Value Handling, Case Study : March 2019 SUSENAS. Jakarta: Politeknik Statistika STIS.753-770.

Nabillah I, Ranggadara I.2020. Mean Absolute Percentage Error untuk Evaluasi Hasil Prediksi Komoditas Laut. *Journal Of Information System.*5(2):250-255. DOI: 10.33633/joins.v5i2.3900.

Sallaby AF, Azlan. 2021. Analysis of Missing Value Imputation Application with K-Nearest Neighbor(K-NN) Algorithm in Dataset. *The International Journal Of Informatics and Computer Science* 5(2):141-144.

Scheaffer RL, Mendenhall III W, Ott RL, Gerow K. 2012. In: *Elementary Survey Sampling Seventh Edition.* Boston.

Wilson DR, Martinez TR.1997. Improved Heterogeneous Distance Functuon. *Journal of Artificial Intelligence Research*6(1): 1-34.