ANALISIS METODE *K NEAREST NEIGHBOR IMPUTATION* (KNNI) UNTUK MENGATASI DATA HILANG PADA ESTIMASI DATA SURVEY

Euis Sartika
Politeknik Negeri Bandung

Abstrak
Ketidaklengkapan suatu data survey (*missing data*) akan menyebabkan inferensi statistik tidak dapat dilakukan. Penghilangan data yang tidak lengkap akan menyebabkan berkurangnya informasi yang dibutuhkan sehingga kurang menggambarkan kondisi yang seenarnya. Terdapat beberapa cara untuk mengatasi missing data tersebut, salah satu adalah dengan mengganti data yang hilang (*missing data*) tersebut. Penelitian ini bertujuan mengkaji analisis k *Nearest Neighbor* sebagai salah satu cara untuk mengimputasi data yang hilang. Metode ini didasarkan pada konsep jarak terdekat dari observasi (k), kemudian pada observasi tersebut diberi pembobotan. Software yang digunakan adalah R versi 3.4.3. Pada penelitian ini digunakan juga metode imputasi yang lain yaitu *Series Mean* dan metode *Algoritma EM* sebagai pembanding. Untuk membandingkan ketiga metode tersebut digunakan nilai RMSE. Hasil menunjukkan bahwa metode *Series Mean* menunjukkan nilai RMSE terkecil. Kelebihan lain dari *k Nearest Neighbor* adalah dapat digunakan untuk imputasi data berskala numerik dan kategorik.

Kata kunci: *missing data*, k *Nearest Neighbor, Series Mean, Algoritma* EM, RMSE

Abstract
The incompleteness of a survey data (missing data) will cause statistical inference can not be done. An incomplete data omission will result in less information needed so as to illustrate the actual condition. There are several ways to overcome the missing data, one is to replace the missing data (missing data). This study aims to examine the analysis of Nearest Neighbor k as one way to imput the missing data. This method is based on the concept of the closest distance from observation (k), then the observation is given weighting. Software used is R version 3.4.3. In this study also used other imputation method that is Series Mean and EM Algorithm method as comparison. To compare the three methods used RMSE value. The results show that the Series Mean method shows the smallest RMSE value. Another advantage of Nearest Neighbor k is that it can be used for numerical and categorical data imputation.

Keywords: *missing data, k Nearest Neighbor, Series Mean, EM Algorithm, RMSE*

I. PENDAHULUAN

Perancangan survei yang baik adalah memilih sampel dengan baik dan mewakili agar kesimpulan terhadap populasi (generalisasi) benar-benar menyimpulkan keadaan populasi yang sebenarnya. Pada kegiatan survei, terkadang tidak semua pertanyaan pada kuisioner diisi dengan lengkap oleh responden , kondisi ini disebut dengan *nonrespon*. Kondisi ini menyebabkan terjadinya *missing data*. *Missing data* (data hilang) dapat mengakibatkan estimasi parameter menjadi kurang akurat karena berkurangnya jumlah atau ukuran data. Kish. L (1965) mendefinisikan *nonrespon* sebagai kegagalan untuk mendapatkan nilai pengamatan dari beberapa unit yang menjadi sampel. *Nonrespon* atau data hilang umumnya dibagi menjadi dua tipe, yaitu *unit nonrespon* dan *item nonrespon*. *Unit nonres*pon terjadi karena unit sampel tidak memberikan respon sama sekali dalam suatu survei. Sedangkan *item nonrespon* dapat terjadi karena direspon oleh responden. Secara umum, *nonrespon* dapat disebabkan karena responden tidak mau menjawab, tidak mampu menjawab atau tidak tahu jawabannya. Longford (2005) menyebutkan bahwa *nonrespon* dapat juga terjadi karena terdapat kesalahan dalam penulisan jawaban atau dalam proses input data. Data *Screening* adalah proses pembersihan data atau persiapan data sebelum melakukan analisis kuantitatif. Jika data hilang (*missing data*) dibiarkan begitu saja maka inferensi statistik untuk data lengkap dengan metode standar tidak dapat dilakukan. Beberapa metode telah banyak dikembangkan untuk memperkecil akibat negatif dari data hilang. Hal-hal yang biasa dilakukan peneliti untuk mengantisipasi data hilang antara lain :

a. jika jumlahnya hanya sebagian kecil, data dibiarkan saja atau tidak dilakukan tindakan apapun;
b. menghapus record data yang tidak lengkap, jadi untuk analisis data hanya menggunakan record yang lengkap namun hal ini dapat mengakibatkan data yang ada tidak dapat menggambarkan kondisi yang sebenarnya;
c. melakukan estimasi untuk melengkapi data yang hilang.

Prosedur penghapusan record data yang tidak lengkap adalah kurang baik, dikarenakan

penghapusan unit-unit pengamatan yang mempunyai data hilang akan mengurangi ukuran sampel yang sudah ditentukan dari awal penelitian. Hal ini tentu saja akan mengurangi ketepatan pendugaan populasi. Levy and Lemeshow(2003) menyebutkan bahwa, jika unit-unit pengamatan yang dihilangkan dalam analisis sangat berbeda dengan unit-unit yang tersisa maka hasil dugaan akan menjadi bias.

Berdasarkan pernyataan yang disebutkan di atas, terdapat beberapa metode untuk mengatasi permasalahan yang ditimbulkan oleh data hilang dalam survei. Namun dalam penelitian ini hanya dikaji metode K *Nearest Neighbor Imputation* yang dianggap sebagai metode yang paling mudah dan populer.

Rumusan Permasalahan adalah sebagai berikut :

1. Bagaimana mengatasi kasus data hilang pada data hasil survei ?
2. Bagaimana penggunaan analisis metode *K Nearest Neighbor Imputation* pada data survey ?

Tujuan penelitian :

1. Mengkaji metode imputasi *K Nearest Neighbor Imputation* untuk mengestimasi data hilang.
2. Menerapkan metode imputasi *Series Mean, Algoritma EM*, dan analisis *K Nearest Neighbor Imputation* pada data hilang (missing data).
3. Membandingkan tingkat ketepatan imputasi data dari metode imputasi *Series Mean, Algoritma EM*, dan analisis *K Nearest Neighbor Imputation*

Peta jalan (road map) penelitian yang telah dilakukan lebih mengarah kepada pemodelan. Dimulai dari pemodelan survival, pemodelan Regresi Logistik Biner, pemodelan Chaid, pemodelan Regresi Logistik Ordinal, dan pemodelan Regresi Berganda. Semuanya ditujukan untuk pengembangan ilmu pengetahuan dan peningkatan teknologi karena penggunaan software statistika di dalamnya. Sedangkan penelitian terkait peningkatan teknologi pada proses pembelajaran yang telah dilakukan adalah penggunaan kalkulator dan software MS Excel dan SPSS untuk peningkatan proses pembelajaran praktikum mata kuliah Statistika Bisnis di Polban.

Pada penelitian ini, dikembangkan metode imputasi yakni metode yang digunakan untuk mengatasi data hilang (*missing data*). Metode ini penting diketahui para peneliti sebelum melakukan analisis data agar tidak terjadi hasil analisis data yang bersifat bias. Jadi, penelitian ini jelas berdasarkan road map pengembangan ilmu pengetahuan dan peningkatan teknologi dalam hal ini penggunaan software statistika, karena secara tidak langsung berkontribusi pada pemecahan masalah secara global yakni pemecahan pada masalah *missing data.* Metode imputasi ini merupakan hal pokok yang harus diketahui para peneliti sebelum melakukan analisis data pemodelan agar mendapatkan hasil atau model yang akurat.

II. TINJAUAN PUSTAKA

Data *screening*

Tujuan utama dari data *screening* ini adalah khususnya menghilangkan data-data yang hilang (*missing* data) atau data-data yang dianggap tidak normal. Misalkan data dari kuesioner yang tidak diisi oleh responden pada beberapa item, mungkin disebabkan karena terlewat, lupa, atau bisa saja responden tidak berkenan mengisinya. Mungkin banyak juga ditemui data-data yang diisikan oleh responden tetapi tidak sesuai dengan apa yang diminta. Proses mempersiapkan data melalui prosedur data *screening* ini dapat banyak membantu peneliti dalam menormalisasi distribusi data dan akan berpengaruh dalam pemilihan alat analisis data untuk uji parametrik atau non parametrik.

Beberapa hal yang harus dilakukan dalam data *screening* :

1. Merespon *Missing Data*
   Missing data adalah data (value) yang terlewat atau hilang dari sebuah isian kuesioner, test, atau instrumen lainnya. *Missing value* selalu berarti hilangnya data secara keseluruhan dengan kata lain hilangnya keseluruhan data responden.
2. Mendeteksi *outlier* (penc*ilan)*
   Outlier adalah data yang berada di luar batas kewajararan. Perlu diketahui, bahwa pada beberapa kasus nilai *outlier* ini bisa saja berupa bilangan desimal diluar nilai 0. Bila dibiarkan begitu saja maka *outlier* ini sangat mempengaruhi distribusi data. Pada sample kecil, *outlier* dengan mudah diidentifikasi. Tetapi, untuk kasus data dengan sampel besar maka diperlukan cara lain untuk mendeteksi *outlier* tersebut. Langkah terakhir setelah merespon data yang hilang dan *outlier* adalah mengganti nilai yang hilang dengan nilai seri rerata seperti yang disediakan oleh SPSS.
3. Transformasi data
   Transformasi data merupakan langkah terakhir dalam merespon data hilang. Field (2013) menyatakan bahwa terdapat empat jenis transformasi data, yaitu: transformasi log, transformasi akar kuadrat, transformasi berbanding terbalik, serta transformasi nilai berlawanan.

Pola data Hilang

Data hilang menurut Enders (2010) dikelompokkan ke dalam 4 tipe, yakni :

1. *Univariat nonrespon* adalah pola data hilang dimana hilangnya data terletak pada satu variabel.
2. *Multivariate nonrespon* adalah pola dimana dalam suatu survei terdapat variabel yang lengkap dan variabel yang tidak lengkap (data hilang), seperti responden menolak untuk menjawab.
3. Data hilang monoton adalah pola data hilang dimana data yang hilang pada pengukuran tertentu selalu hilang pada pengukuran berikutnya. Secara visual, pola monoton menyerupai tangga.
4. Data hilang umum memiliki bentuk yang paling umum dari nilai-nilai yang hilang, artinya memiliki pola yang menyebar pada data secara acak.
5. Data hilang terencana yaitu pola dimana data dengan sengaja dihilangkan, hal ini biasanya dilakukan dengan tujuan mengumpulkan sejumlah besar item kuesioner dan cara ini juga dilakukan dengan maksud mengurangi beban responden.
6. Variabel laten adalah pola yang menarik karena menampilkan nilai-nilai variabel laten yang hilang untuk seluruh sampel.

Mekanisme Data Hilang

Terdapat tiga mekanisme data hilang yaitu:

1) *Missing Completely at Random* (MCAR)
   Misalkan jika data hilang ditentukan oleh sebuah koin, maka mekanisme data hilangnya adalah MCAR.
2) *Missing at Random* (MAR)
   Misalkan pada responden perempuan yang menolak untuk menjawab pertanyaan tentang berat badan atau tinggi badan.
3) *Missing Nonignorable at Random* (MNAR)
   Misalkan seseorang tidak menjawab pertanyaan tentang pendapatan, karena pendapatan yang diperoleh sangat rendah.

Metode Imputasi

Metode Imputasi adalah pengisian nilai untuk data hilang pada suatu survei. Metode imputasi dikelompokkan menjadi dua yaitu :

1. *Single Imputation*
   Salah satu pendekatan dalam prosedur berbasis imputasi adalah *single imputation*. Dalam *single imputation*, data hilang diisi dengan suatu nilai (nilai tunggal) dapat berupa nilai penduga seperti : *mean imputation*, *cold deck imputation*, *hot deck imputation* (Rubin, 2014). Masalah umum yang sering terjadi dalam *single imputation* adalah menempatkan kembali nilai hilang dengan nilai tunggal dan kemudian memperlakukannnya sebagaimana nilai tersebut merupakan nilai sebenarnya (Little dan Rubin, 2014). Hal ini merupakan kelemahan dari *single imputation.*
2. *Multiple imputation* (Imputasi Ganda)
   Dengan keterbatasan *single imputation*, maka selanjutnya dikembangkan metode *multiple imputation* (imputasi ganda). *Multiple imputation* memiliki sejumlah manfaat sebagai suatu pendekatan data hilang. Karena dapat mengisi nilai hilang dengan lebih dari satu kemungkinan, atau sebanyak m kali imputasi. Nilai m dapat berkisar pada 3 sampai dengan 5, atau dengan kata lain imputasi dilakukan maksimal 5 kali (Allison, 2000).

Prosedur untuk Menangani Missing Data

Saat ini terdapat beberapa metode untuk menangani permasalahan *missing data* dalam analisis statistik. Metode-metode tersebut dapat dikelompokkan ke dalam kategori sebagai berikut:

1) Record dengan Unit yang Lengkap (*completely recorded units*)
   Pada kategori ini digunakan pendekatan konsep matriks.
2) Prosedur berbasis Imputasi.
   Imputasi merupakan suatu alternatif yang umum dan fleksibel. Dalam prosedur ini, *missing data* dilengkapi bisa dengan cara menduga langsung atau menggunakan penduga berbasis korelasi. Terdapat beberapa macam pendekatan untuk imputasi jenis ini, antara lain:
   a. *Hot deck imputation*, dimana dari unit-unit yang tercatat disubstitusikan terhadap missing data.
   b. *Cold deck imputation*, dimana *missing data* diganti oleh suatu nilai yang konstan.
   c. *Mean imputation*, yaitu dimana nilai yang hilang diganti oleh rata-rata (mean) dari kelompok sampel unit terkait.
   d. *Regression (correlation) imputation*, yaitu dimana *missing data* dari suatu variabel diestimasi menggunakan nilai penduga dari regresi atau korelasi variabel tersebut pada variabel lainnya yang diketahui.
3) Prosedur *Weighting* (pembobotan), yaitu prosedur mengganti data hilang dengan nilai estimasi yang biasanya didasarkan pada *design weight*, yaitu proporsional secara terbalik terhadap peluang pemilihan sampelnya.
4) Prosedur berbasis Model, yaitu suatu prosedur yang dibentuk dengan menentukan suatu model sebagian data yang hilang (*missing data*) tersebut dan selanjutnya melakukan inferensi berbasis pada *likelihood* di bawah model tersebut. Parameter diestimasi dengan suatu prosedur iteratif *maximum likelihood* dimulai dengan unit atau cases yang lengkap.

Metode *K Nearest Neighbor Imputation* (KNNI)

Metode imputasi KNN merupakan salah satu metode untuk mengatasi data hilang pada data multivariat yang paling mudah dan populer. Kelebihan dari metode imputasi KNN adalah :

a. Metode imputasi KNN dapat digunakan untuk memprediksi dua tipe data , data diskret menggunakan nilai modus dan data kontinu dengan menggunakan dengan nilai mean.
b. Metode imputasi KNN tidak membutuhkan pembentukan model prediksi untuk setiap item yang mengalami *missing data* (Batista, G.E. dan Monard, M.C. 2002.).

Sedangkan kelemahan dari metode imputasi KNN adalah pada saat menentukan pengamatan yang paling sesuai dengan pengamatan yang mempunyai nilai yang hilang, algoritma imputasi KNN akan mencari melalui semua dataset. Kelemahan ini akan berpengaruh apabila dataset yang digunakan cukup besar karena waktu yang dibutuhkan menjadi sangat lama. Tetapi metode imputasi KNN masih tetap merupakan metode yang cukup baik untuk imputasi data yang hilang (Laencina, Gomez, Vidal, dan Verleysen, 2009).

Langkah-langkah pengerjaan imputasi *missing data* dengan metode KNNI adalah sebagai berikut:

1. menentukan nilai K, yaitu banyaknya observasi terdekat yang akan digunakan;
2. menghitung jarak antara observasi yang mengandung missing data pada variabel ke-j dengan observasi lainnya yang tidak mengandung missing data pada variabel yang bersesuaian dengan menggunakan rumus:

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{m} (x_{aj} - x_{bj})^2} \quad \text{.....(1)}$$

Keterangan :

$d(x_a, x_b)$: jarak antar observasi target x_a dan observasi x_b

$x_{aj,}$: nilai pengamatan ke - j pada observasi target x_a, $j = 1,2,..., m$

$x_{bj,}$: nilai pengamatan ke - j pada observasi lainnya x_b, $j = 1,2,..., m$

3. Menentukan nilai K observasi terdekat berdasarkan nilai jarak terkecil. Nilai variabel pada K observasi terdekat ini yang akan digunakan untuk proses imputasi pada observasi yang mengandung nilai missing.
4. Menghitung bobot (*weight*) pada setiap K observasi terdekat. Observasi yang paling dekat akan mendapatkan bobot yang paling besar.
5. Menghitung nilai rata-rata pada K observasi terdekat yang tidak mengandung nilai *missing data* dengan prosedur *weighted mean estimation* yaitu dengan rumus :

$$\hat{x}_j = \frac{1}{KW} \sum_{k=1}^{K} W_k v_{kj} \quad \text{.....(2)}$$

Keterangan :

$W = \sum_{k=1}^{K} W_k$

dengan W_k : bobot observasi tetangga terdekat ke-k yang dirumuskan :

$$W_k = \frac{1}{d(x, v_k)^2} \quad \text{.....(3)}$$

6. Melakukan proses imputasi *missing data* pada observasi yang mengandung *missing data* dengan menggunakan nilai rata-rata yang diperoleh pada tahap 5.

Beberapa penelitian mengenai imputasi data yang telah dilakukan antara lain: Purnaningsih, E (2015) menyatakan bahwa *imputasi hot deck* fraksional mengganti setiap *item nonresponse* dengan sekumpulan nilai imputasi, sedangkan pada *imputasi Nearest Neighbor* setiap item nonresponse diisi oleh item response berdasarkan tetangga terdekat variabel penjelas. Izzah, A dan Hayati, N (2013) menyebutkan bahwa secara rata-rata metode *imputasi KNN-GA* memiliki nilai MSE terendah dan hasil akurasi klasifikasi yang tinggi. Selanjutnya penelitian Siregar, Sartika (2015) menyebutkan bahwa semakin besar missing data yang terjadi pada data produktivitas ubi kayu maka rata-rata nilai RMSE nya juga semakin besar pada setiap nilai K yang digunakan. Artinya bahwa semakin sedikit data yang hilang maka tingkat akurasinya pun semakin baik. Hasil penelitian Mawarsari, U (2016) menyimpulkan bahwa *metode imputasi kNN-GA* dapat membantu dalam memperoleh nilai k optimum dan dapat melakukan seleksi variabel dengan baik sehingga dapat meningkatkan akurasi kNN.

Metode *Series Mean*

Pada metode *Series Mean*, penggantian data yang hilang dilakukan dengan nilai rata-rata (mean) dari data yang ada. Mean diperoleh dari nilai kelompok data dengan menjumlahkan individu dalam kelompok data tersebut dibagi dengan jumlah individu yang terdapat dalam kelompok tersebut. Rumus rata-rata (mean) dinyatakan sebagai berikut :

$$Me = \frac{\sum X_i}{n} \quad \text{.....(4)}$$

Keterangan :

Me = nilai median
X_i = nilai individu i sampai dengan n
n = Banyaknya individu

Algoritma Expectation Maximization (EM)

Salah satu metode optimisasi yang bersifat iteratif untuk estimasi *Maksimum Likelihood* (ML) dalam mengatasi masalah data yang tidak lengkap (*incomplete data*) adalah Algoritma EM. Metode ini dikembangkan pertama kali oleh Demster, Laird, dan Rubin pada tahun 1977. Dalam setiap langkah iterasi pada Algoritma EM mengandung dua tahap, yaitu tahap Ekspektasi (E step) dan tahap Maksimisasi (M step). Algoritma EM ini pada dasarnya hampir sama dengan pendekatan *ad hoc* yakni estimasi missing data. Langkah-langkahnya sebagai berikut :

a) mengisi data yang hilang dengan *nilai estimasi,*
b) mengestimasi parameter,
c) mengestimasi ulang *data yang hilang* tadi dengan parameter baru yang diestimasi,
d) mengestimasi ulang parameter dan terus berulang-ulang sampai diperoleh nilai konvergen terhadap suatu nilai tertentu.

E step* dan *M step

E step bertugas menemukan ekspektasi bersyarat dari data hilang dengan syarat data diketahui nilainya (*observed*) dan penduga parameternya, kemudian mensubstitusikan nilai ekspektasi yang diperoleh terhadap *missing data*. Dalam hal ini *missing data* yang dimaksud bukanlah Y_{miss} tapi fungsi dari Y_{miss} yang muncul dalam complete data loglikelihood, yaitu $\ell(\theta \mid Y)$. Misal $\theta^{(t)}$ adalah penduga parameter θ saat ini dan misal $Y_{miss} = Z$, maka E step pada EM bermaksud mencari ekspektasi loglikelihoodnya jika θ adalah $\theta^{(t)}$:

$$(\theta|\theta^{(t)}) = {}_{(Y_{miss}|Y;\theta(t))}[\ell(\theta;Y)] = \int \ell(\theta;Y) f(Y_{miss}|Y_{obs},\theta=\theta^{(t)})\, dY_{miss} \quad .....(5)$$

M step pada EM bertugas menentukan $\theta^{(t+1)}$ dengan memaksimumkan ekspektasi loglikelihood tersebut. Perhatikan notasi berikut :

$(\theta^{(t+1)} \mid \theta^{(t)}) \geq (\theta \mid \theta^{(t)})$, untuk semua θ

$$\theta^{(t+1)} = \arg\max Q(\theta \mid \theta^{(t)}) \quad .....(6)$$

Langkah-langkah metode Algoritma EM secara garis besar adalah sebagai berikut :

a. E-step : estimasi statistik syarat cukup (*sufficient statistic*) untuk data lengkap Yt dengan terlebih dahulu menghitung nilai ekspektasinya.
b. Nilai θ^n diperoleh M-step dengan menggunakan MLE *(Maximum Likelihood Estimation)* dari Yt
c. Iterasi dianggap selesai jika nilai $\theta^{(t)}$ konvergen, atau nilai $(\theta^{(t+1)} - \theta^{(t)})$ mendekati nol. Hasilnya adalah berupa deret dari nilai-nilai $\theta^{(0)} \rightarrow \theta^{(1)} \rightarrow \ldots$ dimulai dari suatu nilai $\theta^{(0)}$ tertentu.

Kelebihan dan kekurangan Algoritma EM

Beberapa keunggulan Algoritma EM dibanding pendekatan lainnya yaitu:

a. Secara numerik, Algoritma EM lebih stabil sebab dalam setiap iterasinya nilai likelihood-nya naik.
b. Dibawah kondisi umum, algoritma EM konvergen terhadap suatu nilai,. yaitu dimulai dari suatu nilai sembarang $\theta^{(0)}$ akan selalu konvergen terhadap suatu lokal *maximizer*, terkecuali jika salah dalam menentukan nilai awal $\theta^{(0)}$.
c. Algoritma EM cenderung mudah diterapkan, karena didasarkan penghitungan *data lengkap.*
d. Algoritma EM mudah diprogram, karena tidak melibatkan baik integral ataupun turunan dari likelihood.
e. Algoritma EM hanya memerlukan ruang *harddisk* yang sedikit memori di komputer karena tidak menggunakan matriks ataupun invers dalam setiap iterasi.
f. Analisis relatif lebih mudah dibanding metode lain.
g. Dengan memperhatikan kenaikan likelihood secara monoton pada iterasi, maka mudah untuk memonitor konvergensi dan kesalahan program.
h. Bisa digunakan untuk mengestimasi nilai dari *missing data.*

Adapun kelemahan dari Algoritma EM antara lain:

a. Tidak dapat menentukan estimasi matriks kovarian dari penduga parameter.
b. Algoritma EM konvergen secara lambat, jika terlalu banyak *informasi yang tidak lengkap*
c. Algoritma EM tidak menjamin konvergen pada suatu nilai maksimum global jika terdapat nilai maksima lebih banyak.
d. Dalam beberapa masalah, E step mungkin secara analisis akan tidak nyata.

Nilai RSME (*Root Square Mean Error*)

Untuk mengukur tingkat akurasi hasil prakiraan suatu model, digunakan RMSE . RMSE adalah salah metode untuk mengevaluasi teknik peramalan atau mengukur tingkat akurasi hasil prakiraan suatu model. RMSE menyatakan nilai rata-rata dari jumlah kuadrat dari suatu model prakiraan. Nilai RMSE yang kecil memberi petunjuk bahwa variasi (keragaman) nilai yang dihasilkan model prakiraan mendekati variasi nilai observasinya. Seperti yang dikatakan oleh Makridakis, *at. Al*, bahwa salah satu ukuran kesalahan dalam peramalan adalah nilai tengah akar kuadrat atau *Root Mean Square Error* (*RMSE*) yang dinyatakan dalam trumus :

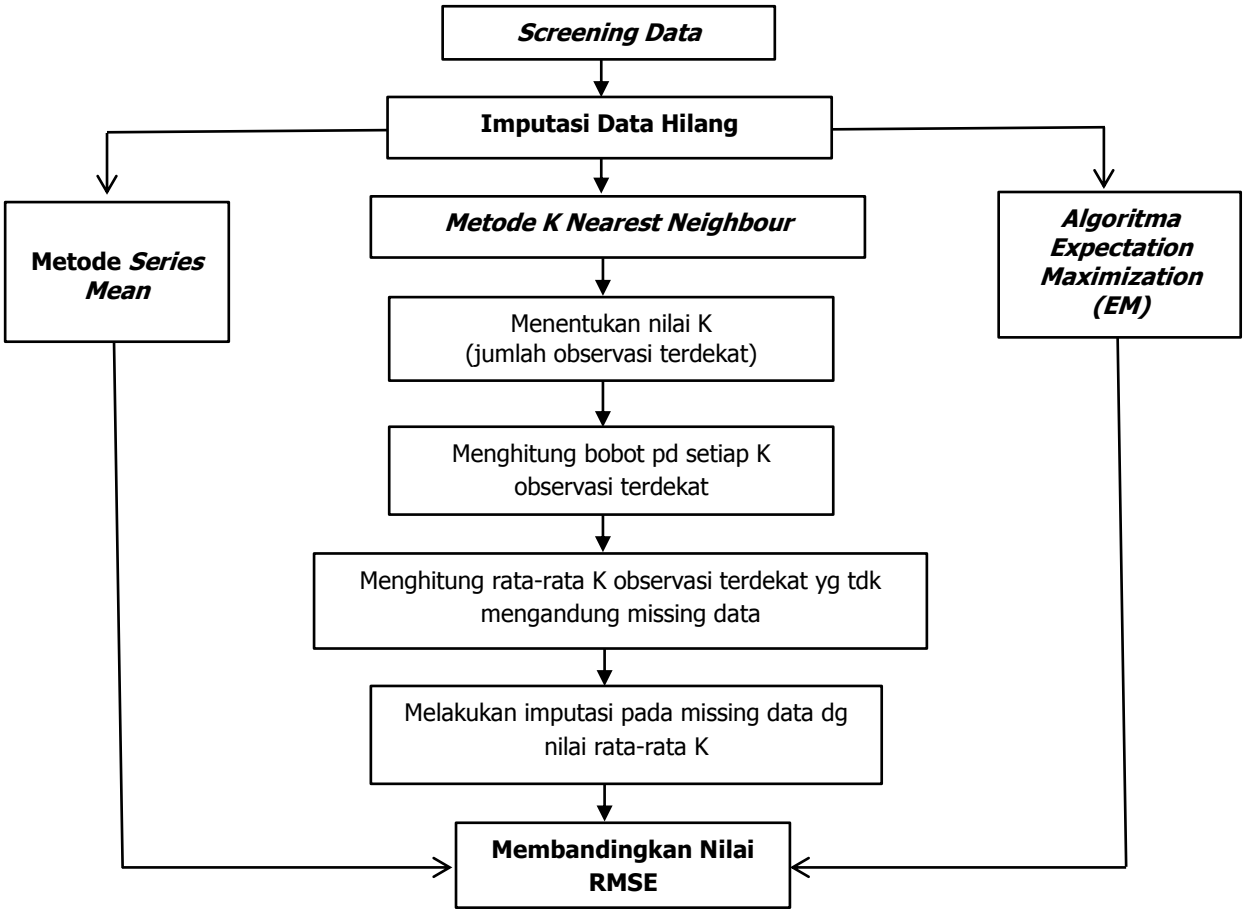
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad .....(7)$$

Keterangan :
n = banyaknya kelompok dat
$\hat{y}_i$: Nilairamalan ke - i
y_i : Nilai observasi ke-i

III. METODE PENELITIAN

Metode Imputasi untuk mengatasi data hilang yang dikaji adalah *Metode K Nearest Neigbor Imputation* menggunakan software statistik R. Data normal yang dibangkitkan secara acak dalam bentuk matriks berukuran 10 x 10 (100 buah). Data terdiri dari 10 variabel. Sebagai pembanding akan digunakan metode *Series Mean* dan *Algoritma Expectation Maximization (EM) Technique* yang diolah dengan software SPSS. Metode *K Nearest Neigbor Imputation* yang ada dalam SPSS adalah diperuntukkan untuk masalah klasifikasi.

Untuk lebih jelasnya langkah-langkah penelitian digambarkan dalam diagram berikut :



Gambar 1. Diagram alir langkah-langkah penelitian

IV. HASIL DAN PEMBAHASAN

Tabel 1. Data acak normal sembarang

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
-0,5909136	-1,2698175	0,5556009	-0,1327224	-0,8325065	0,71664	0,5802716	-1,0669137	0,486708	0,5824858
-1,2725507		0,7261897	0,29699	0,2969556	-0,4461283		-1,0532693	-0,3403039	-1,1041681
0,5547398	0,4748735		-0,9493917		-0,9937785	-2,8199645			-0,035263
-0,4779765	0,1647818	0,6167311	-0,5149731	0,5240514	-0,4602781	-0,5508051	-1,0620243	-0,5138356	0,3161907
-1,0876783		0,6659499	-0,7223724	-0,9512409	-1,6054705		-0,3696951	0,3554919	0,3202675
-0,0615328	0,9486815	-0,5464601	0,1544475		-0,8225022	-0,2504348	-1,0389393	0,078109	0,5251606
-0,8253603	-0,2906253	-0,3028428	-0,847321	0,7985286	-0,0975193	-0,4166732			-1,1586756
-1,1536619			-1,5900281		0,2903993	-0,0689836	-1,0736161	-1,0373835	-1,0373835
-1,4950447	-0,5419533	0,5766574	-1,2412777	-1,4089572	-0,7106984	0,3005158		0,3141792	-1,4155193
-0,3593544	-0,2622265	0,4048126	-2,0869817	0,2682486	0,1690456	-0,6818003	0,5829092	0,5829092	-1,0197956

Berdasarkan tabel 1, dapat diperlihatkan bahwa jumlah dua buah data hilang . Sedangkan variabel V1, V4, V6, dan V10, tidak ada data yang hilang. Secara total hilang untuk masing-masing data hilang mencapai 15 % atau sekitar 15 buah.

Perbandingan Statistika Deskriptif (summary) dari ketiga metode

Tabel 2. *Summary* Metode *Series Mean*

Statistik	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Mean	-0,67693334	-0,11089798	0,33707983	-0,76336306	-0,1864172	-0,396029	-0,4884843	-0,7259355	-0,0092657	-0,40267005
Median	-0,70813695	-0,110898	0,48020675	-0,7848467	-0,1864172	-0,4532032	-0,4525788	-0,8824374	0,03442165	-0,5275293
Mode	-1,4950447a	-0,110898	0,3370798	-2,0869817a	-0,1864172	-1,6054705a	-0,4884843	-0,7259355	-0,0092657	-1,4155193a
Std. Deviation	0,622330253	0,589344421	0,426568638	0,755219591	0,700178139	0,6851781	0,91162983	0,51588156	0,50146046	0,808221766
Variance	0,387	0,347	0,182	0,57	0,49	0,469	0,831	0,266	0,251	0,653
Range	2,0497845	2,218499	1,2726498	2,3839717	2,2074858	2,3221105	3,4002361	1,6565253	1,6202927	1,9980051
Minimum	-1,4950447	-1,2698175	-0,5464601	-2,0869817	-1,4089572	-1,6054705	-2,8199645	-1,0736161	-1,0373835	-1,4155193
Maximum	0,5547398	0,9486815	0,7261897	0,29699	0,7985286	0,71664	0,5802716	0,5829092	0,5829092	0,5824858
Sum	-6,7693334	-1,1089798	3,3707983	-7,6336306	-1,864172	-3,9602904	-4,8848427	-7,2593551	-0,0926571	-4,0267005

Tabel 3. *Summary* Metode *Algoritma EM*

Statistik	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Mean	-0,6769333	-0,1881513	0,4589146	-0,763363	0,163161	-0,39603	-0,41392	-0,48096	0,095279	-0,40267005
Median	-0,708137	-0,2619236	0,5661292	-0,784847	0,282602	-0,4532	-0,31503	-0,83885	0,259004	-0,5275293
Mode	-1,4950447a	-1,2698175a	-,5464601a	-2,0869817	-1,408957	-1,605470	-2,819964	-1,073616	-1,037383	-1,4155193a
Std. Deviation	0,62233025	0,61680931	0,5826348	0,7552196	1,402685	0,685178	0,933412	0,802111	0,566871	0,80822177
Variance	0,387	0,38	0,339	0,57	1,968	0,469	0,871	0,643	0,321	0,653
Range	2,0497845	2,218499	2,1344981	2,3839717	4,871212	2,322111	3,400236	2,302562	1,860467	1,9980051
Minimum	-1,4950447	-1,2698175	-0,54646	-2,086982	-1,40896	-1,60547	-2,81996	-1,07362	-1,03738	-1,4155193
Maximum	0,5547398	0,9486815	1,588038	0,29699	3,462255	0,71664	0,580272	1,228946	0,823083	0,5824858
Sum	-6,7693334	-1,8815128	4,5891464	-7,633631	1,631612	-3,96029	-4,13924	-4,80964	0,952787	-4,0267005

Tabel 4. *Summary* Metode *k Nearest Neighbor*

Statistik	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Mean	-0,6769333	-0,3093869	0,22015627	-0,7633631	-0,2265908	-0,396029	-0,4959106	-0,44942085	-0,4175065	-0,0601429
Median	-0,7081369	-0,3345931	0,53572535	-0,7848467	0,2712407	-0,4532032	-0,4499088	-0,7043172	-0,7043172	0,01598785
Mode	-1,4950447a	-1,2698175a	-1,0109258a	-2,0869817a	-1,5187722a	-1,6054705a	-2,8199645a	-1,0736161a	-1,0736161a	-1,0373835a
Std. Deviation	0,62233025	0,69489108	0,61010144	0,75521959	0,85663068	0,68517807	0,91198404	0,739863274	0,76776054	0,52295438
Variance	0,387	0,483	0,372	0,57	0,734	0,469	0,832	0,547	0,589	0,273
Range	2,0497845	2,218499	1,7371155	2,3839717	2,3173008	2,3221105	3,4002361	1,9545702	1,9545702	1,6202927
Minimum	-1,4950447	-1,2698175	-1,0109258	-2,0869817	-1,5187722	-1,6054705	-2,8199645	-1,0736161	-1,0736161	-1,0373835
Maximum	0,5547398	0,9486815	0,7261897	0,29699	0,7985286	0,71664	0,5802716	0,8809541	0,8809541	0,5829092
Sum	-6,7693334	-3,0938687	2,2015627	-7,6336306	-2,2659077	-3,9602905	-4,9591062	-4,4942085	-4,1750647	-0,6014291

Perbandingan Nilai RMSE dari ketiga Metode

Tabel 5. Nilai RMSE dari Ketiga Metode

Metode	RMSE
Series Mean	0,492620544
Algoritma EM	1,392974874
k.Nearest Neighbor	0,81432119

Berdasarkan tabel 5 dapat ditunjukkan metode *Series Mean* dianggap metode paling tepat dalam imputasi data karena nilai RMSE-nya paling kecil.

IV. HASIL PENGUJIAN

Pengujian ANOVA

ANOVA yang digunakan adalah satu arah (*one way*), karena hanya faktor nilai imputasi sebagai variabel dependen sedangkan yang menjadi yang menjadi faktor adalah ketiga metode imputasi.

Tabel 6. Hasil uji Anova

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	,534	2	,267	,441	,643
Within Groups	179,567	297	,605		
Total	180,101	299			

Berdasarkan tabel 6, disimpulkan bahwa Imputasi data hilang dari ketiga metode , yakni : *Series Means, Algoritma EM,* dan *k Nearest Neighbor* tidak memberikan
Hal ini terbukti dari nilai P-value yang diberikan yakni 0,643 > 0,05

V. KESIMPULAN DAN SARAN

Kesimpulan

1. Untuk mengatasi kasus data hilang (*missing value*) dapat digunakan metode antara lain, metode *Series Mean*, metode *Algoritma EM*, dan metode *k nearest neighbor*.
2. Proses pengimputan, metode *Series Mean*, proses pengimputan datanya didasarkan pada nilai rata-rata dari data yang ada, pengimputan data hilang dengan metode Algoritma EM didasarkan pada konsep optimisasi yang bersifat iteratif. Konsep dasar dari metode k *Nearest Neighbor* adalah pengimputan data hilang menggunakan jarak antara observasi terdekat
3. Berdasarkan nilai imputasi yang dilakukan oleh ketiga metode di atas, pengujian ANOVA menunjukkan bahwa secara statistik tidak ada perbedaan yang signifikan.
4. Nilai RMSE metode Series Mean menunjukkan nilai yang paling kecil, artinya metode Series Mean menghasilkan nilai imputasi data yang paling tepat.

Saran

1. Diharapkan ada penelitian lain dengan menggunakan metode-metode Imputasi data yang berbeda.
2. Penelitian lainnya diharapkan mengandung data hilang dengan persentase data hilang yang bervariasi.
3. Pada metode k *Nearest Neighbor*, diharapkan nilai k atau banyaknya observasi terdekat juga dicobakan berbeda-beda sehingga dapat dihasilkan imputasi data yang lebih akurat.

DAFTAR PUSTAKA

Basuki, R. (2009). *Imputasi Berganda Menggunakan Metode Regresi dan MetodePredictive Mean Matching untuk Menangani Missing Data* (Doctoral dissertation, Tesis. Institut Teknologi Sepuluh November: Surabaya)

Batista, G.E. dan Monard, M.C. 2002. A Study of K-Nearest Neighbour as an Imputation Method. Second International Conference on Hybrid Intelligence Systems, 87, 251-260.

Dempster, A.P., N. M. Laird, and D.B., Rubin, 1977; *Maximum Likelihood from Incomplite Data via EM Algorithm (with discucion)*, Journal of the Royal Statistical Society B, Vol. 39, No.1-38, pp.25.

Evriyanto. Yudi,. (2004), " *Perbandingan Metode Imputasi Untuk Mengestimasi Data Hilang Pada Data Kesehatan Ibu Dan Anak Di Jawa Timur*",.Tesis .Universitas Airlangga.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Hendrawati,T. 2015, "Kajian Metode Imputasi Dalam Menangani Missing Data"Staf Pengajar Statistika Universitas PadjadjaranProsiding Seminar Nasional Matematika dan Pendidikan Matematika UMS 2015.

Izzah, A. I. (2013). Imputasi Missing data menggunakan Metode K-Nearest Neighbour Dengan Optimasi Algoritma Genetika. *Melek IT Information Technology Journal*, *2*(2).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning (p. 68).* New York: Springer.

Kish.L,.1965, "Survey Sampling" Key: citeulike:553273. Posts Export Citation.

Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: methods and applications*. John Wiley & Sons.

Laencina, P.J.G., Gomez, J-L.S., Vidal, A.R.F., dan Verleysen, M. 2009. K Nearest Neighbours with Mutual Information for Simultaneous Classification and Missing Data Imputation. Neurocomputing, 72, 1483-1493. Elsevier.

Leech, N.L., Barret, K.C., & Morgan, G.A. (2005). SPSS for intermediate statistics: Use and interpretation (2nd ed). London: Lawrence Erlabaum Associates.

Longford, N. T. (2005). Model selection and efficiency—is 'Which model…?'the right question?. *Journal of the Royal Statistical*

Society: Series A (Statistics in Society), *168*(3), 469-472.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Makridakis, *at. Al, (1998)*, *Forecasting : Methods and Applications*. Third Edition. New York : John Wiley & Sons, inc

Makridakis, S. et al., (1982) "*The Accuracy of Extrapolative (Time Series Methods): Results of a Forecasting Competition"*, Journal of Forecasting, Vol. 1, No. 2, pp. 111-153 (lead article

Morgan, G.A., Leech, N.L., Gloeckner, G.W., & Barret, K.C. (2004). SPSS for introductory statistics: Use and interpretation (2nd ed.). London: Lawrence Erlabaum Associates.

Mawarsari, U. (2016). Imputasi Missing Data Dengan K-Nearest Neighbor Danalgoritma Genetika. *AdMathEdu*, *6*(1).

Purnaningsih, E., & Haryatmi, S. (2015). *Imputasi Hot Deck Fraksional Dan Imputasi Nearest Neighbour Untuk Mengatasi Item Nonresponse* (Doctoral Dissertation, Universitas Gadjah Mada).

Siregar, S. Y., Toharudin, T., & Tantular, B. (2015). *Performa Metode K Nearest Neighbor Imputation (KNNI) Untuk Menangani Multivariate Missing Data* (Doctoral dissertation, Universitas Padjadjaran).

Wilcox, R.R. (2005). Introduction to robust estimation and hypothesis testing (2nd ed.). Elsevier.

https://myenglish01.wordpress.com/2014/07/31/data-screening-membersiapkan-data-untuk-analisa-kuantitatif/ ..... diaksesKamis, 3 Agustus 2017

https://statistikakomputasi.wordpress.com/2010/04/08/analisis-missing-data-menggunakan-algoritma-em-2/..diunduh pada tanggal 28 November 2017