PERBANDINGAN METODE IMPUTASI: METODE *MEAN* DAN METODE *K NEAREST NEIGHBOR* (KNN) UNTUK MENGATASI DATA HILANG PADA DATA SURVEI

(Skripsi)

Oleh

DESY NUR FITRIANA MURJITO

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2021

ABSRTACT

THE COMPARISON IMPUTATION METHOD : MEAN METHOD AND K NEAREST NEIGHBOR (KNN) METHOD TO OVERCOME THE MISSING DATA OF SURVEY DATA

By

DESY NUR FITRIANA MURJITO

One of the problems in survey is often experienced is there are some units that do not respond to some of the questions so that it makes data are incomplete or data are missing. Imputation method is one of the ways to overcome the missing data. Mean imputation and K Nearest Neighbor Imputation are two method that can be used to this research. The purpose of the research is to compare the imputation method to estimate the missing data with Mean Imputation and K Nearest Neighbor (KNN) Imputation and search which method is better among two methods. Based on simulation study with 1000 replication KNN Imputation method has a smaller the average value of Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE) than Mean Imputation. Therefore, it is concluded that KNN Imputation method is better than Mean Imputation Method.

Keywords : Missing Data, Imputation, Mean Imputation, K Nearest Neighbor Imputation, Mean Square Error, Mean Absolute Percentage Error

ABSRTAK

PERBANDINGAN METODE IMPUTASI: METODE *MEAN* DAN METODE *K NEAREST NEIGHBOR* (KNN) UNTUK MENGATASI DATA HILANG PADA DATA SURVEI

Oleh

DESY NUR FITRIANA MURJITO

Salah satu masalah yang sering dialami di dalam survei adalah ditemukannya unit-unit yang tidak merespon sejumlah pertanyaan yang diajukan, sehingga menyebabkan data tidak lengkap atau data hilang. Salah satu cara mengatasi data hilang adalah metode imputasi. Dalam penelitian ini digunakan metode Imputasi *Mean* dan Imputasi *K Nearest Neighbor*. Tujuan penelitian ini adalah membandingkan metode imputasi untuk mengestimasi data hilang dengan metode Imputasi *Mean Imputation* dan Imputasi *K Nearest Neighbor* (KNN) dan mencari metode mana yang lebih baik di antara kedua metode tersebut. Berdasarkan hasil studi simulasi dengan 1000 kali ulangan diperoleh bahwa metode Imputasi KNN menghasilkan rata-rata nilai Mean Square Error (MSE) dan Mean Absolute Percentage Error (MAPE) yang lebih kecil dibandingkan metode Imputasi *Mean*. Dengan demikian, maka.dapat disimpulkan bahwa metode imputasi KNN lebih baik daripada metode imputasi *Mean*.

Kata Kunci : Data Hilang, Imputasi, *Mean Imputation*, *K Nearest Neighbor Imputation*, *Mean Square Error*, *Mean Absolute Percentage Error*

PERBANDINGAN METODE IMPUTASI: METODE *MEAN* DAN METODE *K NEAREST NEIGHBOR* (KNN) UNTUK MENGATASI DATA HILANG PADA DATA SURVEI

Oleh

DESY NUR FITRIANA MURJITO

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2021

Judul Skripsi	:	**PERBANDINGAN METODE IMPUTASI: METODE *MEAN* DAN METODE *K NEAREST NEIGHBOR* (KNN) UNTUK MENGATASI DATA HILANG PADA DATA SURVEI**
Nama Mahasiswa	:	Desy Nur Fitriana Murjito
Nomor Pokok Mahasiswa	:	1717031042
Jurusan	:	Matematika
Fakultas	:	Matematika dan Ilmu Pengetahuan Alam

Menyetujui

1. Komisi Pembimbing

Dr. Khoirin Nisa, S.Si., M.Si.
NIP. 19740726 200003 2 001

Dra. Dorrah Aziz, M.Si.
NIP.19610128 198811 2 001

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman S.Si. M.Si.
NIP. 19740316 200501 1 001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Khoirin Nisa, S.Si., M.Si.** .....

Sekretaris : **Dra. Dorrah Aziz, M.Si** .....

Penguji
Bukan Pembimbing : **Prof. Drs. Mustofa, M.A., Ph.D.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Suripto Dwi Yuwono, M.T.
NIP. 19740705 200003 1 001

Tanggal Lulus Ujian Skripsi: **10 Agustus 2021**

PERNYATAAN SKRIPSI MAHASISWA

Saya yang bertanda tangan di bawah ini :

Nama	: **Desy Nur Fitriana Murjito**
Nomor Pokok Mahasiswa	: **1717031042**
Judul	: **PERBANDINGAN METODE IMPUTASI: METODE *MEAN* DAN METODE *K NEAREST NEIGHBOR* (KNN) UNTUK MENGATASI DATA HILANG PADA DATA SURVEI**
Jurusan	: **Matematika**

dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri dan semua tulisan dalam skripsi ini telah mengikuti penulisan karya ilmiah Universitas Lampung.

Bandar Lampung, 10 Agustus 2021

Desy Nur Fitriana Murjito
NPM. 1717031042

RIWAYAT HIDUP

Penulis merupakan putri sulung dari dua bersaudara yang dilahirkan pada 24 Desember 2000 di kota sang Ibu dibesarkan, Boyolali.

Penulis menyelesaikan sekolah dasar di SD Negeri 1 Mojolegi pada 2011. Studi penulis dilanjutkan di SMP Muhammadiyah 3 Bandar Lampung yang ditamatkan pada 2014. Pada 2017 penulis lulus dari SMA Negeri 13 Bandar Lampung dan pada tahun yang sama dinyatakan resmi menjadi salah satu mahasiswa baru jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam melalui ujian Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN).

Selama berstatus sebagai mahasiswa, penulis berkesempatan menjadi anggota magang Biro Sirkulasi dan Periklanan Natural 2017, Kepala Biro Kesekretariatan Natural 2018, serta menjadi anggota dari komunitas Sahabat Difabel Lampung (Sadila) pada 2019. Pada Januari s.d. Februari 2020 penulis mengikuti Kuliah Kerja Nyata (KKN) di Tulang Bawang dan pada Juli s.d. Agustus 2020 penulis melakukan kerja praktik di UPTD Laboratorium Dinas Bina Marga dan Bina Konstruksi Provinsi Lampung.

PERSEMBAHAN

Puji dan syukur ke hadirat Allah SWT atas limpahan rahmat-Nya sehingga penulis dapat menyelesaikan karya ilmiah ini dengan baik.

Dengan segala kerendahan hati Penulis mempersembahkan karya ini kepada :

Ayahanda dan Ibunda tercinta, yang selalu memberikan cinta, kasih sayang, pengorbanan, motivasi, nasihat, dan mendoakan penulis di setiap waktu untuk keberhasilan penulis.

Dosen-dosen Pembimbing dan Pembahas yang selalu memberikan motivasi dan sosok yang sangat berjasa dibalik penulisan karya ilmiah ini.

Sahabat-sahabat terkasih, terimakasih atas kebersamaan, doa, dan waktu yang telah kalian berikan untuk penulis.

Almamater yang kubanggakan, Universitas Lampung.

KATA INSPIRASI

"Sesungguhnya Allah tidak akan mengubah keadaan suatu kaum hingga mereka mengubah diri mereka sendiri."

(Q.S. Ar-Ra'd: 11)

"Allah tidak membebani seseorang melainkan sesuai dengan kesanggupanya."

(Q.S. Al-baqarah : 286)

"Don't put to much pressure on yourself."

(Ten Lee)

SANWACANA

Puji dan syukur ke hadirat Allah SWT atas limpahan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Perbandingan Metode Imputasi: Metode *Mean* dan Metode *K Nearest Neighbor* (KNN) untuk Mengatasi Data Hilang pada Data Survei". Shalawat serta salam tak lupa kepada Rasulullah SAW yang telah menjadi suri tauladan yang baik bagi umat beliau. Penulis menyadari bahwa dalam peulisan skripsi ini tak lepas dari bimbingan, bantuan, kerja sama, serta dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis hendak mengucapkan terima kasih kepada :

1. Ibu Dr. Khoirin Nisa, S.Si., M.Si., selaku Dosen Pembimbing I, yang selalu membimbing, memberikan ide, arahan, kritik dan saran, serta motivasi kepada penulis selama proses pembuatan skripsi.
2. Ibu Dr. Dorrah Aziz, S.Si., M.Si., selaku Dosen Pembimbing II, yang telah membimbing, memberikan saran, solusi serta pembelajaran kepada penulis dalam proses menyelesaikan skripsi.
3. Bapak Prof. Drs. Mustofa Usman, M.A., Ph.D., selaku Dosen Pembahas, yang telah memberikan evaluasi dan saran dalam perbaikan skripsi.
4. Bapak Dr. Aang Nuryaman, S.Si., M.Si., selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
5. Bapak Dr. Eng. Suripto Dwi Yuwono, M.T. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
6. Seluruh Dosen, Staf, dan Karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

7. Ayahanda, Ibunda, dan keluarga tercinta yang senantiasa memberi dukungan, doa, nasihat, serta mendampingi penulis dalam proses menyelesaikan skripsi.
8. Sahabat-sahabat terkasih: Belmut, Kis, dan Padew yang bersedia menemani hari-hari penulis, memberi pendapat, serta kekuatan dalam menyelesaikan skripsI, *you all guys, are my nice sisters.*
9. Teman-teman Matematika 2017 yang telah memberikan pengalaman berharga bagi penulis.
10. WayV gege yang telah memberikan motivasi dan menghibur penulis dalam proses menyelesaikan skripsi.
11. Seluruh pihak yang selalu memberikan dukungan dan doa kepada penulis, baik dalam penyusunan skripsi maupun kehidupan sehari-hari, yang tidak dapat disebutkan satu persatu.
12. Penulis yang tetap kuat dan mampu bertahan, tidak menyerah, serta berjuang selangkah demi selangkah hingga tahap ini, terima kasih.

Bandar Lampung, 10 Agustus 2021
Penulis,

Desy Nur Fitriana Murjito

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Analisis data dengan statistika sudah sangat familiar di kalangan peneliti dan akademisi, namun terkadang terdapat salah satu masalah yang sering terjadi yaitu adanya beberapa data hilang atau data tidak lengkap. Saat melakukan sensus ataupun survei, salah satu masalah yang sering dialami adalah ditemukannya unit-unit yang tidak merespon sejumlah pertanyaan yang diajukan (*nonrespon*). Menurut Kish (1965), *nonrespon* didefinisikan sebagai kegagalan untuk mendapatkan nilai pengamatan dari beberapa unit yang menjadi contoh. *Nonrespon* dalam beberapa referensi sering disebut dengan data hilang. Data hilang umumnya dibagi menjadi dua tipe, yaitu *unit nonrespon* dan *item nonrespon*. *Unit nonrespon* terjadi karena adanya unit contoh yang tidak memberikan respon sama sekali dalam survei. Sedangkan *item nonrespon* terjadi karena terdapat item dalam kuesioner yang tidak direspon oleh responden. Secara umum, *nonrespon* disebabkan karena kesalahan sistem seperti tidak adanya respon terhadap sensor atau perangkat penerima *input*, dapat pula disebabkan karena kesalahan manusia seperti kelalaian dalam pengumpulan data, ketidakmampuan responden dalam memberikan jawaban yang akurat, atau responden tidak berkenan memberikan jawaban yang akurat (Izzah & Hayatin, 2013).

Sensus sebuah populasi merupakan usaha yang dilakukan untuk mendapatkan informasi dari semua unit dalam popolasi yang dijadikan objek penelitian, sedangkan survei hanya dilakukan terhadap beberapa unit populasi (contoh). Perancangan survei

yang baik akan memilih contoh dengan benar agar kesimpulan terhadap populasi yang dijadikan objek penelitian bersifat terandal dan representatif untuk menyimpulkan keadaan populasi.

Adanya masalah *nonrespon* atau data hilang menimbulkan data hasil survei atau sensus tidak lengkap. Hilangnya data akan membawa kesulitan bagi peneliti, karena dengan adanya data hilang maka data tidak dapat dianalisis dengan baik. Akibatnya, peneliti perlu metode statistik yang bisa dipertanggungjawabkan secara ilmiah untuk mengatasi data hilang, sehingga meskipun ada data yang hilang tetap dapat dianalisis dengan baik.

Terdapat beberapa cara yang dapat dilakukan untuk mengatasi data hilang diantaranya imputasi tunggal (*single imputation*) dan imputasi ganda (*multiple imputation*). Metode imputasi tunggal yang paling umum digunakan adalah imputasi dengan imputasi rata-rata (*mean imputation*). Pengembangan dari metode imputasi tunggal adalah metode imputasi ganda (Wilsen, dkk., 2018). Salah satu metode imputasi ganda, yaitu metode *K-Nearest Neighbor Imputation* (KNNI).
Beberapa penelitian yang mengkaji tentang data hilang antara lain Susanti tahun 2018 tentang *K Nearest Neighbor Imputation* dalam Imputasi *Missing Data*, Euis tahun 2018 tentang Analisis Metode *K Nearest Neighbor Imputation* (KNNI) untuk Mengatasi Data Hilang pada Estimasi Data Survey, serta Nisa, dkk. tahun 2020 tentang *Analysis of Variance for Strip Plot Design ith Missing Values: Bias Correction of the Mean Squares*.

Berdasarkan uraian di atas, maka dalam penelitian ini penulis akan mengkaji perbandingan metode Imputasi *Mean* dan Imputasi KNN untuk pendugaan data hilang.

1.2 Tujuan Permasalahan

Berdasarkan latar belakang di atas, tujuan penelitian ini adalah untuk mengkaji dan membandingkan metode *Mean Imputation* dan metode *K Nearest Neighbor Imputation* (KNNI).

1.3 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah :

1. Menambah wawasan pengetahuan kepada penulis khususnya tentang menduga data hilang secara simulasi dengan metode *Mean Imputation* dan metode KNNI.
2. Memberikan informasi sebagai bahan dalam menambah referensi tentang menduga data hilang secara simulasi dengan metode *Mean Imputation* dan metode KNNI.
3. Sebagai bahan tinjauan pustaka untuk setiap pihak yang memerlukan.

II. TINJAUAN PUSTAKA

2.1 Pengertian Survei

Survei dapat diartikan sebagai inspeksi, pemeriksaan, penilikan, dan peninjauan pada suatu objek penelitian. Sedangkan pengertian sampel adalah himpunan bagian dari populasi yang dipilih penelitian untuk diteliti. Dari definisi survei dan sampel tersebut dapat disimpulkan bahwa survei sampel merupakan salah satu metode pengumpulan data melalui sebagian unit dalam populasi dan hasilnya merupakan nilai-nilai perkiraan (estimasi).

2.2 Penarikan Sampel Acak Sederhana

Sampel acak sederhana diperoleh dengan menggunakan metode penarikan sampel sederhana (*simple random sampling*). Cara penarikan *simple random sampling* dapat melalui dua cara (Supranto, 2009).

1. Pemilihan *simple random samping* tanpa pengembalian, yaitu metode pemilihan sampel untuk unit yang sudah terpilih tidak dikembalikan dalam sampel (*without replacement*).
2. Pemilihan *simple acak random sampling* dengan pengembalian, yaitu metode pemilihan sampel untuk unit yang sudah terpilih dikembalikan dalam sampel agar dapat dipilih kembali (*with replacement*).

2.3 Data Hilang

Literatur statistik tentang masalah data hilang pertama kali diperkenalkan oleh Orchard dan Woodbury (1972). Data hilang merupakan adanya ketidaklengkapan informasi atau data pada suatu variabel. Adanya data hilang menyebabkan metode baku untuk data lengkap tidak dapat digunakan untuk menganalisis data.

Pada penerapannya, metode analisis untuk data lengkap sering digunakan untuk data-data yang mempunyai data hilang dengan cara menghilangkan unit pengamatan yang mempunya data hilang. Terdapat beberapa alasan logis yang memperlihatkan kenyataan bahwa prosedur tersebut tidak baik. Pertama, penghapusan unit-unit pengamatan yang mempunyai data hilang akan mengurangi contoh yang telah ditentukan sejak awal penelitian. Hal tersebut otomatis akan mengurangi ketepatan pendugaan populasi. Kedua, jika unit-unit pengamatan dihilangkan dalam analisis yang sangat berbeda dengan unit-unit yang tersisa, maka hasil dugaan akan menjadi berbias (Levy & Lemeshow, 1999).

2.3.1 Pola Data Hilang

Menurut Buuren (2012), ada beberapa pola data hilang, yaitu :

1. Univariat dan Multivariat
   Data hilang dikatakan memiliki pola univariat apabila hanya ada satu variabel yang mengalami masalah data hilang.
2. Monoton dan *non-monoton* (umum)
   Data hilang dikatakan berpola monoton ketika data yang hilang pada pengukuran tertentu selalu hilang pada pengukuran berikutnya. Ketika pola monoton tidak terpenuhi, data hilang disebut *non-monoton* (umum).
3. Terhubung dan tidak terhubung
   Pola terhubung terjadi apabila data hasil observasi dapat diakses dari observasi yang lainnya dengan cara berpindah secara vertikal atau horizontal. Ketika

antardata hasil observasi tidak dapat dihubungkan, baik dengan perpindahan vertikal atau horizontal, data hilang dikatakan memiliki pola tidak terhubung.

2.3.2 Jenis Data Hilang

Menurut Harlan (2016), data hilang dibedakan menjadi dua, yaitu :

1. Data hilang terencana (*planed missing data*) terjadi karena data direncanakan dengan sengaja oleh peneliti sesuai desain penelitian. Hal ini biasanya terjadi karena jumlah pertanyaan dalam kuesioner terlalu banyak dan pengumpulan data tertentu menggunakan pemeriksaan dengan biaya yang sangat mahal.
2. Data hilang tak terencana (*unplanned missing data*) terjadi karena ketidakberhasilan mendapatkan data dari subjek penelitian, data tercatat sengaja dihapus karena nilainya berada di luar rentang kewajaran dan data yang tidak dapat ditelusuri lagi penyebab kekosonganya.

2.3.3 Mekanisme Data Hilang

Menurut Buuren (2012), terdapat mekanisme data hilang diklasifikasikan sebagai berikut :

1. *Missing Completely at Random* (MCAR)
   Mekanisme data hilang secara MCAR tidak memiliki keterkaitan (saling bebas), baik dengan variabel yang diamati ataupun yang tidak diamati. Oleh karena itu, mekanisme data hilang secara MCAR terjadi apabila besarnya peluang suatu data akan hilang adalah sama dan acak
2. *Missing at Random* (MAR)
   Data yang hilang dengan mekanisme MAR tidak selalu terjadi secara acak seperti pada mekanisme MCAR, melainkan bergantung pada data hasil observasi. Hal tersebut dapat dilihat berdasarkan nilai peluang suatu data menjadi hilang. Nilai peluang suatu data akan hilang dalam mekanisme MAR tidak memiliki bobot

yang sama, melainkan bergantung dengan hasil pengukuran dari observasi lain yang diteliti.

3. *Missing Nonignorable at Random* (MNAR)
   Mekanisme MNAR berbeda dengan mekanisme MCAR ataupun MAR. Proses hilangnya suatu data tidak hanya bergantung dengan data yang telah diobservasi, tetapi bergantung juga dengan berbagai faktor di luar pengukuran yang dilakukan.

2.3.4 Prosedur untuk Mengatasi Data Hilang

Menurut Sartika (2018), 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 imptation*, 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)

   Prosedur Weighting merupakan 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

   Prosedur berbasis Model merupakan 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.

2.4 Nilai Koefsien Korelasi

Korelasi dalam teori probabilitas dan statistika juga disebut koefisien korelasi yaitu nilai yang menunukkan kekuatan dan arah hubungan linier antara dua peubah acak (*random variable*). Koefisen korelasi (r) digunakan untuk mengetahui kuat atau lemahnya hubungan antara variabel-variabel bebas dan variabel terikat. Nilai koefisien korelasi memiliki nilai dari negatif satu hingga satu $(-1 \leq r \leq 1)$. Pada variabel-variabel yang memiliki nilai koefisien $> 0{,}5$ atau $< -0{,}5$ dikatakan memiliki korelasi yang kuat. Jika nilai kofisien korelasi bernilai positif maka kenaikan / penurunan nilai variabel bebas pada umumnya diikuti oleh kenaikan / penurunan nilai variabel terikat, sedangkan untuk variabel-variabel yang memiliki nilai koefisien korelasi bernilai negative negatif berarti kenaikan / penurunan nilai variabel bebas diikuti oleh penurunan / kenaikan variabel terikat.

Berikut merupakan tabel interval koefisien korelasi.

Tabel 1. Interval Koefiien Korelasi

Interval Koefisien (r)	**Tingkat Hubungan**
$\mathbf{0 \leq r < 0,2}$	Sangat Rendah
$\mathbf{0,2 \leq r < 0,4}$	Rendah
$\mathbf{0,4 \leq r < 0,6}$	Cukup Kuat
$\mathbf{0,6 \leq r < 0,8}$	Kuat
$\mathbf{0,8 \leq r \leq 1,0}$	Sangat Kuat

Adapun rumus perhitungan untuk menentukan nilai koefisien korelasi (r) antara variabel terikat Y terhadap variabel bebas X dengan n jumlah data ditunjukkan pada persamaan berikut (Gasperz, 1992). Rumus ini juga biasa disebut koefisien korelasi Pearson (*Pearson's product coefficient of correlattion*).

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i-1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_1^2 - (\sum_{i=1}^{n} X_i^2)^2}\sqrt{\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i^2)^2}}. \quad (1)$$

2.5 Metode Imputasi

Menurut Sartika (2018), metode Imputasi adalah pengisian nilai untuk data hilang pada suatu survei. Metode imputasi dikelompokkan menjadi dua, yaitu :

1. Imputasi Tunggal (*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* (Little & 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. Imputasi Ganda (*Multiple imputation*)
   Dengan keterbatasan *single imputation*, maka selanjutnya dikembangkan metode *multiple imputatio*n. 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.

2.5.1 *Mean Imputation*

Metode *mean imputation* merupakan salah satu metode imputasi yang paling umum digunakan. Imputasi dengan metode *mean* mengisi data hilang dalam suatu variabel dengan rata-rata dari semua nilai yang diketahui pada suatu variabel (Acuna & Rodriguez, 2004). Terdapat kelemahan pada imputasi dengan metode *mean* yaitu mengurangi *varian* pada variabel. Hal tersebut disebabkan karena data hilang pada data diisi dengan nilai yang sama. Rumus untuk metode *mean imputation* adalah :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

dengan n adalah banyaknya data lengkap.

2.5.2 *K Nearest Neighbor Imputation (KNNI)*

Menurut Susanti, dkk. (2018), *Nearest Neighbor* (NN) adalah sebuah metode yang menggunakan algoritma *supervised learning*. *Supervised learning* bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Terdapat dua jenis algoritma NN, yaitu INN dan KNN. INN

(*Nearest Neighbor*) merupakan pendekatan yang melakukan klasifikasi pada satu data terdekat, sedangkan KNN (*K Nearest Neighbor*) merupakan pendekatan yang melakukan klasifikasi pada K data terdekat, dengan $K > 1$.

KNN merupakan metode yang digunakan untuk melakukan klasifikasi terhadap objek yang diamati berdasarkan beberapa data yang jaraknya paling dekat dengan objek yang tersebut. Pada klasifikasi, KNN bekerja dengan menghitung jarak antara data baru (*data testing*) dengan data yang sudah diketahui kelasnya (*data training*) menggunakan jarak *euclidian* (Susanti, dkk., 2018).

Kelebihan dari metode imputasi KNN adalah sebagai berikut :

1. Metode imputasi KNN dapat digunakan untuk memprediksi dua tipe data, data diskret mengguankan nilai modus dan data kontinu dengan mengguanakan nilai rata-rata.
2. Metode imputasi KNN tidak membutuhkan pembentukan model prediksi untuk setiap item yang mengalami kehilangan data (Batista & Monard, 2002).

Sedangkan untuk kelemahan dari metode imputasi KNN adalah pada saat menentukan pengamatan yang paling sesuai dengan pengamatan yang memiliki nilai yang hilang, algoritma imputasi KNN akan mencari melalui seluruh dataset. Kelemahan ini akan berpengaruh apabila dataset yang diamati cukup besar karena waktu yang dibutuhkan menjadi sangat lama (Sartika, 2018). Tetapi metode imputasi KNN masih tetap merupakan metode yangcukup baik untuk mengimputasi data hilang (Laencina, dkk., 2009).

Menurut Susanti, dkk. (2018), penanganan *missing data* dengan KNN diawali dengan menentukan sejumlah tetangga terdekat atau observasi terdekat yang disimbolkan dengan K, kemudian menghitung jarak terkecil dari setiap observasi yang tidak mengandung *missing data*. Tahapan pengerjaan imputasi data hilang dengan metode KNNI dapat dijelaskan sebagai berikut :

1. Menentukan parameter K, K adalah jumlah observasi terdekat yang akan digunakan.
2. Menghitung jarak antara observasi yang mengandung data hilang pada variabel ke- j dengan observasi lainnya yang bersesuaian dengan rumus jarak *euclidian*, yaitu (Larose, 2005) :

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{m} \left(x_{aj} - x_{bj}\right)^2} \quad (3)$$

$d(x_a, x_b)$ adalah jarak antara observasi yang mengandung data hilang dengan observasi yang tidak mengandung data hilang, x_{aj} adalah nilai dari variabel ke-j pada setiap observasi yang mengandung missing data dengan $j = 1,2, \dots, m$, x_{bj} adalah nilai dari variabel lainnya pada setiap observasi yang tidak mengandung data hilang dengan $j = 1,2, \dots, m$.

3. Mengurutkan jarak berdasarkan observasi yang memiliki nilai jarak terbesar hingga observasi yang memiliki nilai jarak terkecil.
4. Menentukan K observasi terdekat berdasarkan jarak terkecil.
5. Melakukan imputasi data hilang dengan menghitung nilai *weight mean estimation* pada K observasi terdekat yang tidak mengandung nilai data hilang dengan rumus sebagai berikut (Larose, 2005) :

$$\bar{x}_j = \frac{\sum_{k=1}^{K} w_k v_k}{\sum_{k=1}^{K} w_k} \quad (4)$$

$\bar{x}_j$ adalah estimasi rata-rata berbobot, v_k adalah nilai pada data lengkap pada variabel yang mengandung data hilang berdasarkan observasi dari k, K adalah jumlah observasi terdekat yang digunakan, k adalah observasi dari K, w_k adalah bobot observasi tetangga terdekat ke-K dengan rumus $w_k = \frac{1}{d(x_{ak}x_{bk})^2}$, dengan $d(x_{ak}x_{bk})$ adalah jarak observasi K.

2.6 *Mean Square Error* (MSE) dan *Mean Absolute Percentage Error* (MAPE)

Nilai MSE dan MAPE dari hasil imputasi data hilang menunjukkan perbedaan hasil hasil prediksi dengan data aktual (Susanti dkk., 2018). MSE merupakan besaran kesalahan hasil prediksi dengan mengkuadratkan masing-masing kesalahan, sehingga semakin kecil nilai MSE maka semakin kecil kesalahan hasil prediksi. MSE dihitung dengan rumus berikut (Makridakis, dkk., 1999) :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - F_i)^2 \tag{5}$$

dimana $X_i$adalah data aktual untuk periode ke-i, F_i adalah hasil prediksi untuk periode ke-i dan n adalah jumlah periode waktu. Sedangkan MAPE merupakan persentase ukuran kesalahan dari hasil prediksi. Semakin kecil nilai MAPE maka akan semakin kecil kesalahan hasil prediksi, begitupun sebaliknya jika semakin besar nilai MAPE maka akan semakin besar kesalahan hasil prediksi. Hasil imputasi sangat baik jika nilai MAPE $< 10\%$, sedangkan hasil imputasi baik jika nilai MAPE dianatara 10% dan 20% (Susanti, dkk., 2018). MAPE dapat dihitung dengan rumus berikut (Makridakis, dkk., 1999) :

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{X_i - F_i}{X_i}\right|}{n}100\% \tag{6}$$

dimana X_i adalah data aktual untuk periode ke-i, F_i adalah hasil prediksi untuk periode ke-i, dan n adalah jumlah periode waktu.

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

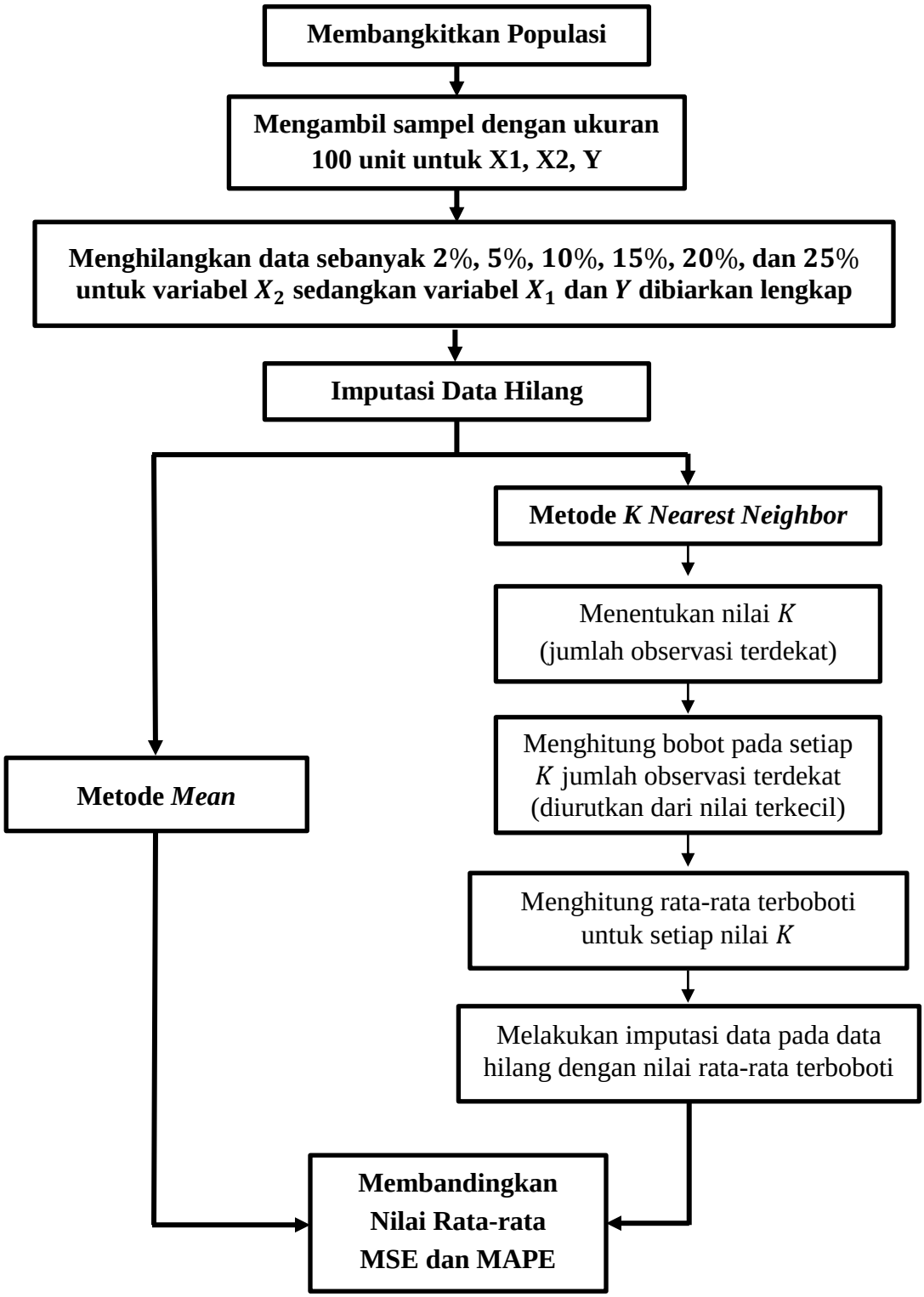
Penelitian ini dilakukan pada semester ganjil tahun ajaran 2020/2021 di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini menggunakan data hasil simulasi dengan bantuan *software R*. Mengambil sampel sebanyak 100 unit dari data hasil bangkitan populasi sebanyak 1000 unit untuk masing-masing variabel X_1, variabel X_2,dan variabel Y dengan syarat variabel X_1 dan variabel X_2 memengaruhi variabel Y. Diasumsikan variabel ini merupakan peubah yang berpeluang besar terjadi *nonrespon* karena beberapa sebab dalam survei yang dilakukan. Adapun beberapa variabel yang dianggap memengaruhi hasil Tes Potensi Akademik (TPA) tersebut adalah usia dan nilai *Intelligence Quotient* (IQ). Sehingga, dalam pembangkitan data, ketiga variabel, usia (X_1), nilai IQ (X_2), dan hasil tes TPA (Y), tersebut akan dibuat agar memiliki korelasi yang cukup tinggi. Masing-masing variabel dibangkitkan dari sebaran normal.

3.3 Metode Penelitian

1. Membangkitkan populasi sebesar 1000 unit dibangkitkan.
2. Dari data populasi diambil contoh berukuran 100 untuk masing-masing variabel, kemudian terhadap data dilakukan penghilangan data yang berbeda-beda.
3. Perlakuan penghilangan data hanya diberikan kepada peubah X_2 sedangkan X_1 dan Y dibiarkan lengkap.
4. Simulasi data hilang pada peubah X_2 dengan *simple random sampling* tanpa ulangan dengan presentasi data hilang sebanyak 2%, 5%, 10%, 15%, 20%, dan 25%.
5. Menduga data hilang dengan metode *Mean Imputation* dan metode KNNI. Setiap gugus data diimputasi menggunakan metode KNN dengan nilai tetangga terdekat k yang dicobakan adalah 5, 10, 15, 20, dan 30.
6. Menghitung nilai MSE dan MAPE dari kedua metode imputasi.
7. Mengulangi Langkah 4 hingga 5 sebanyak 1000 kali untuk memberikan peluang yang sama kepada setiap data.
8. Menghitung rata-rata MSE dan MAPE dari 1000 ulangan untuk setiap nilai k yang dicobakan pada metode KNNI.
9. Membandingkan rata-rata nilai MSE dan MAPE dari hasil imputasi variabel X_2.



Gambar 1. Diagram Alir Metode Penelitian

V. KESIMPULAN

Berdasarkan hasil penelitian Perbandingan Metode Imputasi Metode *Mean* dan Metode *K Nearest Neighbor* (KNN) untuk Mengatasai Data Hilang pada Survei dapat disimpulkan bahwa hasil imputasi data hilang menggunakan metode KNN lebih baik daripada metode *mean* karena rata-rata nilai MSE dan MAPE dari imputasi metode KNN lebih kecil daripada metode *mean*. Nilai K pada hasil imputasi data hilang dipengaruhi oleh besarnya persentase data hilang, semakin besar persentase data hilang maka akan semakin besar rata-rata nilai MSE dan MAPE pada setiap jumlah tetangga terdekat K.

DAFTAR PUSTAKA

Acuna, E. & Rodriguez, C. 2004. The Treatment of Missing Values and Its Effect is The Classifier Accuracy, hlm. 639-647. Proceedings of the Meeting of the International Federation of Classification Societies (IFCS). Illinois of Technologi, Chicago.

Allison, P.D. 2000. Multiple Imputation for Missing: Data A Cautionary Tale. *Sociological Method & Research*. **28**(3): 301-309.

Batista, G. E. & Monard, M. C. 2002. A Study of K-Nearest Neighbour as an Imputation Method, hlm. 251-260. Proceedings of the Second International Conference on Hybrid Intellegence Systems. Santiago, Chile.

Buuren, S.V. 2012. *Flexible Imputation of Missing Data*. USA: CRS Press.

Gasperz, V. 1992. *Teknik Analisis dalam Penelitian Percobaan*. Bandung: Tarsito.

Harlan, J. 2016. *Data Kosong dan Imputasi Ganda*. Depok: Gramedia.

Izzah, A. & Hayatin, N. 2013. *Imputasi Missing Data Menggunakan Algoritma Pengelompokan Data K-Harmonic Means*. Seminar Nasional Matematika dan Aplikasinya.

Kish. 1995. *Survey Sampling*. New York: Willey & Sons.

Laencina, G., Gomez, S., Vidal, F., & Verleysen, M. 2009. K Nearest Neighbours with Mutual Information for Simultaneous Classification and Missing Data. *Neurocomputing*. **72**(7-9): 1483-1493.

Larose, T.D. 2005. *Discovering Knowledge in Data*. New Jersey: Willey & Sons.

Levy, P.S. & Lemeshow. 1999. *Sampling of Population: Methods & Application 3rd ed*. New York: Willey & Sons.

Little, R.J. & Rubin, D.B. 2014. *Statistical Analysis with Missing Data*. New York: Willey & Sons.

Makridakis, S., Wheelwright, S.C., & Mcgee, V.E. 1999. *Metode dan Aplikasi Peramalan*. Jakarta: Binapura Aksara.

Nisa, K., Hamsyiah, N., Usman, M., & Warsono. 2020. Analysis of Variance for Strip Plot Design ith Missing Values: Bias Correction of the Mean Squares. *Journal of Physics*: Conference Series, 1524 (2020), 012049.

Sartika, E. 2018. Analisis K-Nearest Neighbor Imputation (KNNI) untuk Mengatasi Data Hilang pada Estimasi Data Survey. *TEDC*. **12**(3): 219-227.

Supranto, J. 2009. *Statistika Teori dan Aplikasi*. Jakarta: Erlangga.

Susanti, Martha, S. & Sulistianingsih, E. 2018. K Nearest Neighbor *dalam Imputasi* Missing Data. *Buletin Ilmiah Math, Stat, dan Terapannya (Bimaster)*. **7**(1): 9-14.

Wilsen, Rahayu, W., Santi, V.M. 2018. Penerapan Imputasi Ganda dengan Metode Predictive Mean Matching (PMM) untuk Pendugaan Data Hilang pada Model Regresi Linear. *Jurnal Statistika dan Aplikasinya (JSA)*. **1**(1):12-20.