Analisis Perbandingan Metode Imputasi Pada Data Hilang Menggunakan Mean, Median dan Modus dengan Pendekatan Metode Mean Absolute Error (MAE)

Rahma Neliyana[1)], Eka Fidiya Putri[2)], Izza Lutfia[3)], Dea Mutia Risani[4)], Dinda Nababan[5)]

Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera

Email : rahma.122450036@student.itera.ac.id[1)], eka.122450045@student.itera.ac.id [2)], izza.122450090@student.itera.ac.id [3)], dea.122450099@student.itera.ac.id [4)], dinda.122450120@student.itera.ac.id [5)]

1.Pendahuluan

Dalam analisis data, keberadaan nilai yang kosong atau hilang seringkali menjadi tantangan yang harus diatasi karena dapat menjadi faktor yang mempengaruhi hasil analisis (R & D, n.d.). Maka dari itu, menangani nilai-nilai yang hilang merupakan suatu aspek penting untuk diatasi dengan tepat. Data bersih bisa berdiri karena berbagai alasan, sebagai salah aglomerasi informasi, petaka teknis, atau ketidakhadiran pelapor bagian dalam survei. Penanganan informasi bersih menjabat penting karena bisa menakluki mutu dan keakuratan polemik yang dilakukan. Salah tunggal penghampiran kepada informasi bersih adalah memperhatikan petunjuk imputasi, di mana pandangan hidup yang bersih digantikan tambah pandangan hidup estimasi.

Dalam hal ini, penelitian bertujuan untuk melakukan perbandingan antara tiga metode imputasi atau pengisian nilai data yang hilang dengan tiga metode imputasi, yaitu metode mean, median, dan modus, dengan menggunakan pendekatan Mean Absolute Error (MAE). MAE adalah salah satu metode yang digunakan untuk mengukur tingkat keakuratan model peramalan. Nilai MAE menunjukan rata-rata kesalahan(*error)* absolut antara hasil peramalan/prediksi dengan nilai riil (Subagyo & Pangestu, 1986).

2. Metode

2.1 Teknik Imputasi

Teknik imputasi adalah cara mengisi missing value dengan nilai yang adil dari metode imputasi yang digunakan sebelum menjadi data lengkap yang siap modelkan dan dianalisis. Keunggulan dari teknik imputasi adalah kegiatan imputasi data dalam menangani missing value tidak bergantung pada pemilihan metode prediksi dan klasifikasi akan tetapi dapat memilih algoritma pembelajaran yang sesuai setelah imputasi (C. Zhang et al, 2007).

2.2 Imputasi dengan Metode Mean

Mean merupakan salah satu metode imputasi yang paling umum digunakan. Imputasi dengan metode Mean mengisi missing data dalam suatu variabel dengan rata-rata dari semua nilai yang diketahui pada suatu variabel (Rodrigues et al, 2004). Imputasi dengan metode Mean memiliki kelemahan yaitu mengurangi varians pada variabel, karena nilai yang diisikan adalah sama untuk setiap variabel (Graham, 2014).

2.3 Modus

Modus adalah nilai-nilai dalam sebuah kumpulan data yang paling sering muncul atau memiliki frekuensi tertinggi. Dalam statistik, modus digunakan untuk mengidentifikasi nilai yang paling umum atau dominan dalam data.

2.4 Median

Median adalah nilai tengah dari kumpulan data yang telah diurutkan, Dalam statistik, median adalah salah satu ukuran pemusatan data yang penting. Untuk menemukan median, data diurutkan dari nilai terkecil hingga nilai terbesar, kemudian nilai tengahnya diambil sebagai median.

2.5 *Closure*

Closure adalah sebuah fungsi yang dapat disimpan dalam variabel dan biasa dimanfaatkan untuk mewakili (enclose) sebuah proses pada blok tertentu. Variabel yang menyimpan closure memiliki sifat seperti fungsi yang disimpannya. Dengan menggunakan closure, fungsi dapat mengembalikan objek (atau fungsi lain) dan mengingat lingkungan tempat fungsi tersebut diinisiasi. Nilai yang ada pada outer scope masih diingat meskipun fungsi tersebut telah dihapus. Closure memiliki peran yang sangat penting dalam decorator. Closure digunakan untuk menyimpan nilai fungsi yang telah di dekorasi.

2.6 *Mean Absolute Error* (MAE)

Mean Absolute Error (MAE) adalah nilai mutlak dari selisih antara nilai prakiraan dengan nilai sebenarnya (Draxler et al, 2014). Untuk evaluasi model peramalan, MAE lebih intuitif dalam memberikan rata-rata error dari keseluruhan data (Suryanto, 2019). Rumus matematis dari MAE ditunjukkan pada Persamaan 1.

$$MAE = \frac{\sum_{n=1}^{N} \left| \check{r}_n - r_n \right|}{N} \quad (1)$$

dengan
MAE = mean absolute error
Ř*n* = prediction rating
rn = true rating in testing data set

3. Pembahasan

3.1. Deskripsi Dataset

Pada analisis kali ini, kami menggunakan dataset hubungan antara impor India dengan negara Andorra. Dataset ini diperoleh melalui situs *kaggle*. Dimana pada data tersebut terdapat nilai data yang hilang yang nantinya akan dianalisis menggunakan metode imputasi. Berikut merupakan gambar tabel data yang akan digunakan.

Country	Import
Andorra	0.07
Andorra	
Andorra	
Andorra	
Andorra	0
Andorra	0
Andorra	0.02
Andorra	
Andorra	0.03
Andorra	0.01
Andorra	0.09
Andorra	0
Andorra	0.01
Andorra	0.03
Andorra	0.01
Andorra	0
Andorra	5.28
Andorra	0
Andorra	0.09
Andorra	
Andorra	0.04
Andorra	0.03
Andorra	0.01
Andorra	
Andorra	

Gambar 1. Tabel Dataset Impor India dari Negara Andorra

3.2. Kode dan Hasil Pemrograman

Berikut merupakan kode dan hasil pemrograman yang telah kami buat. Secara keseluruhan kode ini menjelaskan tentang pengisian data hilang menggunakan metode imputasi dengan menggunakan mean, median, dan modus. Dimana, nantinya ketiga metode akan dibandingkan melalui pengukuran hasil mean absolute error(MAE).

```
import pandas as pd
import math
from collections import Counter
from sklearn.metrics import mean_absolute_error
```

Gambar 2. Import *Library* dan Modul

Gambar 2 menjelaskan mengenai pengimportan *library* dan modul yang akan digunakan dalam *Python*. Dalam kode tersebut menggunakan *library pandas* dan beberapa modul. Dimana *library pandas* digunakan dalam melakukan manipulasi dan analisis data, modul math digunakan untuk operasi matematika, *Counte*r dari *collections* digunakan untuk menghitung frekuensi kemunculan elemen, serta *mean_absolute_error* dari *sklearn.metrics* digunakan untuk menghitung nilai rata-rata dari selisih absolut antara prediksi dan nilai yang diamati.

```
def isNaN(num):
    return num != num
```

Gambar 3. Membuat Fungsi isNan

Gambar 3 menjelaskan mengenai pembuatan fungsi isNan. Dimana fungsi tersebut bertujuan untuk memeriksa apakah suatu nilai adalah NaN (Not a Number) dalam Python.

```
def data_imputer(strategy='mean'):
    def impute(data):
        if strategy == 'mean':
            mean_val = sum([x for x in data if not isNaN(x) ]) / len([x for x in data if not isNaN(x)])
            return [x if not isNaN(x) else math.ceil(mean_val) for x in data]
        elif strategy == 'median':
            sorted_data = sorted(x for x in data if not isNaN(x))
            median_val = sorted_data[len(sorted_data) // 2]
            return [x if not isNaN(x) else math.ceil(median_val) for x in data]
        elif strategy == 'mode':
            mode_val = Counter(x for x in data if not isNaN(x)).most_common(1)[0][0]
            return [x if not isNaN(x) else math.ceil(mode_val) for x in data]
        else:
            raise ValueError("Unsupported imputation strategy")

    return impute
```

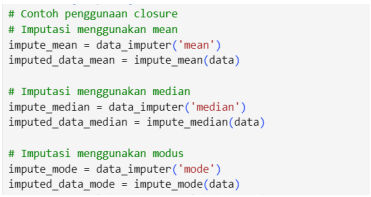
Gambar 4. Fungsi Closure

Gambar 4 menjelaskan mengenai sebuah fungsi yang dibuat di dalam fungsi. Kode tersebut menggunakan konsep *closure* dalam *Python*. Kode tersebut mendefinisikan sebuah fungsi bernama data_imputer yang memiliki parameter optional strategy dengan nilai default 'mean'. Fungsi ini mengembalikan fungsi dalam bentuk closure yang disebut impute. Fungsi impute akan mengambil data sebagai argumen dan melakukan imputasi (penggantian nilai yang hilang) berdasarkan strategy yang ditentukan. Jika strategy adalah 'mean', maka nilai yang hilang akan diganti dengan nilai rata-rata dari data non-hilang. Jika strategy adalah 'median', nilai yang hilang akan diganti dengan median dari data non-hilang. Jika strategy adalah 'mode', nilai yang hilang akan diganti dengan modus dari data non-hilang. Jika strategy tidak memenuhi ketiga kondisi tersebut, kode akan memunculkan ValueError. Lalu fungsi data_imputer akan mengembalikan fungsi impute yang kemudian dapat digunakan untuk melakukan imputasi data sesuai dengan strategi yang dipilih.

```
filename = input("Masukkan nama file (file.csv): ")
data = pd.read_csv(filename)
data= data["Import"]
```

Gambar5. Fungsi Membaca File

Gambar 5 menjelaskan mengenai fungsi yang dibuat untuk membaca file csv yang akan di input. Dalam kode tersebut, file akan dibaca menggunakan *library pandas*. Setelah itu, *library pandas* juga akan menyimpan file dalam variabel data. Kemudian, kode akan memilih kolom *"Import"* dari data yang telah dibaca dan menyimpannya kembali dalam variabel data. Dengan demikian, variabel data hanya akan berisi kolom *"Import"* dari file CSV yang telah dimasukkan *users*.



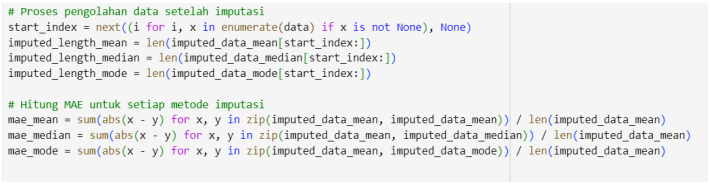
```
# Contoh penggunaan closure
# Imputasi menggunakan mean
impute_mean = data_imputer('mean')
imputed_data_mean = impute_mean(data)

# Imputasi menggunakan median
impute_median = data_imputer('median')
imputed_data_median = impute_median(data)

# Imputasi menggunakan modus
impute_mode = data_imputer('mode')
imputed_data_mode = impute_mode(data)
```

Gambar 6. Fungsi Imputasi Data

Gambar 6 menjelaskan tentang fungsi imputasi data yang menggunakan nilai rata-rata(mean), median, atau modus. Dalam kode tersebut digunakan untuk membuat data imputer yang dapat disesuaikan dengan tipe imputasi yang diinginkan seperti Mean, Median, Modus.



```
# Proses pengolahan data setelah imputasi
start_index = next((i for i, x in enumerate(data) if x is not None), None)
imputed_length_mean = len(imputed_data_mean[start_index:])
imputed_length_median = len(imputed_data_median[start_index:])
imputed_length_mode = len(imputed_data_mode[start_index:])

# Hitung MAE untuk setiap metode imputasi
mae_mean = sum(abs(x - y) for x, y in zip(imputed_data_mean, imputed_data_mean)) / len(imputed_data_mean)
mae_median = sum(abs(x - y) for x, y in zip(imputed_data_mean, imputed_data_median)) / len(imputed_data_mean)
mae_mode = sum(abs(x - y) for x, y in zip(imputed_data_mean, imputed_data_mode)) / len(imputed_data_mean)
```

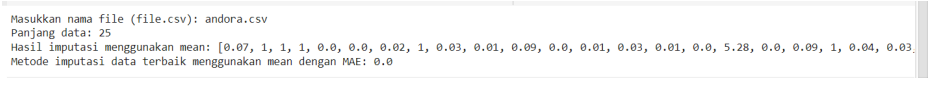
Gambar 7. Fungsi pengolahan data

Gambar 7 menjelaskan Fungsi tersebut digunakan untuk melakukan proses pengolahan data setelah dilakukan imputasi, serta untuk menghitung *Mean Absolut Error (MAE)* untuk metode imputasi yang dilakukan. Kode tersebut digunakan untuk menentukan data yang sudah diimputasi.

```
# Pilih hasil imputasi dengan RMSE terkecil sebagai hasil terbaik
hasil_imputasi_terbaik = ''
if mae_mean <= mae_median and mae_mean <= mae_mode:
    hasil_imputasi_terbaik = 'Mean'
    print("Panjang data:", imputed_length_mean)
    print("Hasil imputasi menggunakan mean:", imputed_data_mean[start_index:])
    print(f"Metode imputasi data terbaik menggunakan mean dengan MAE: {mae_mean}")
elif mae_median <= mae_mean and mae_median <= mae_mode:
    hasil_imputasi_terbaik = 'Median'
    print("Panjang data:", imputed_length_median)
    print("Hasil imputasi menggunakan median:", imputed_data_median[start_index:])
    print(f"Metode imputasi data terbaik menggunakan median dengan MAE: {mae_median}")
else:
    hasil_imputasi_terbaik = 'Modus'
    print("Panjang data:", imputed_length_mode)
    print("Hasil imputasi menggunakan modus:", imputed_data_mode[start_index:])
    print(f"Metode imputasi data terbaik menggunakan modus dengan MAE: {mae_mode}")
```

Gambar 8. Fungsi memilih imputasi

Gambar 8 menjelaskan fungsi memilih imputasi terbaik berdasarkan *Mean Absolute Error (MAE)* terkecil. Kode tersebut digunakan juga untuk membandingkan MAE dari masing-masing metode untuk menentukan metode mana yang memberikan hasil imputasi paling baik. Selain itu fungsi tersebut menyimpan informasi tentang metode imputasi yang memberikan hasil terbaik berdasarkan MAE.



```
Masukkan nama file (file.csv): andora.csv
Panjang data: 25
Hasil imputasi menggunakan mean: [0.07, 1, 1, 1, 0.0, 0.0, 0.02, 1, 0.03, 0.01, 0.09, 0.0, 0.01, 0.03, 0.01, 0.0, 5.28, 0.0, 0.09, 1, 0.04, 0.03
Metode imputasi data terbaik menggunakan mean dengan MAE: 0.0
```

Gambar 9. Hasil kode

Gambar 9 menjelaskan hasil dari kode yang sudah dibuat. Pada gambar tersebut dapat dilihat bahwa dataset dengan kolom import memiliki panjang sebesar 25. Data tersebut juga menampilkan hasil imputasi pada data menggunakan mean. Dari hasil tersebut dapat dilihat bahwa imputasi data terbaik menggunakan MAE diperoleh melalui metode mean.

4. Kesimpulan

Dataset yang digunakan adalah data hubungan impor antara India dan negara Andorra, yang mengandung nilai data yang hilang (NaN) yang akan diimputasi menggunakan metode tertentu. Tiga metode imputasi yang dibandingkan adalah menggunakan mean, median, dan modus. Setiap metode memiliki keunikan dalam pengisian nilai yang hilang berdasarkan karakteristik data tersebut. Tahap-tahap dalam pemrograman meliputi pengimportan library dan modul yang diperlukan, pembuatan fungsi untuk memeriksa nilai NaN, pembuatan fungsi imputasi dengan berbagai strategi (mean, median, modus), serta proses pengolahan data dan perhitungan Mean Absolute Error (MAE) untuk masing-masing metode. Dengan demikian, dapat disimpulkan bahwa metode imputasi menggunakan mean memberikan hasil yang paling baik dalam mengisi nilai yang hilang pada dataset "andora.csv".

Referensi

R, L., & D, R. (n.d.). Statistical Analysis with Missing Data, Third Edition. John Willey & Sons.

Subagyo, & Pangestu. (1986). Forecasting Konsep dan Aplikasi. *Yogyakarta, BPPE UGM.*

Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang, 2007, Semiparametric optimization for missing data imputation,Appl. Intell., vol. 27, no. 1, pp. 79–88, doi:10.1007/s10489-006-0032-0.

Acuna E, Rodrigues C. The Treatment of Missing Values and its Effect is the Classifier Accuracy. Proceedings of the Meeting of the International Federation of Classification Sociaties

(IFCS). 2004 juli 15.

Graham JW. Missing Data Analysis and Design [online]. USA: Springer; 2012 [cited 2014 Des 16]. Available from: Bookfi.org