Mengatasi Data Hilang menggunakan Fungsi Closure Data Imputer pada Dataset Curah Hujan Provinsi Lampung periode 2022

Cyntia Kristina Sidauruk (122450023)[1], Patricia Leondrea Diajeng Putri (122450050)[2], Berliana Enda Putri (122450065)[3] , Feryadi Yulius(122450087)[4], Naufal Fakhri (122450089)[5]

Program Studi Sains Data Institut Teknologi Sumatera
Jl. Terusan Ryacudu, Way Huwi, Kec. Jatiagung, Kabupaten Lampung Selatan, Lampung 35365

Email:[1]cyntia.122450023@student.itera.ac.id, [2]patricia.122450050@student.itera.ac.id, [3]berliana.122450065@student.itera.ac.id, [4]feryadi.122450087@student.itera.ac.id, [5]naufal.122450089@student.itera.ac.id

Abstrak

Persoalan data hilang pada dataset sering menjadi hambatan yang signifikan dalam pemrosesan data untuk mengatasi duduk perkara ini, salah satu pendekatan yang bisa digunakan adalah menggunakan memakai teknik imputasi data. Dalam konteks ini, kami melakukan pendekatan menggunakan fungsi closure pada Python buat melakukan imputasi data dengan cara metode statistik mirip mean, median, serta mode. Fungsi closure memungkinkan kita untuk membuat sebuah fungsi yang dapat diadaptasi serta dapat dipergunakan kembali dengan parameter-parameter yg tidak selaras sesuai kebutuhan. dengan memakai fungsi closure, kita bisa dengan simpel diterapkan pada berbagai metode imputasi data menggunakan satu fungsi yang dapat diadaptasi.hasil percobaan menunjukkan telah berhasil menyelesaikan problem data yg hilang pada dataset cuaca provinsi Lampung dengan akurasi memuaskan. Dapat disimpulkan, pendekatan ini memiliki potensi untuk meningkatkan kualitas serta kegunaan aneka macam dataset yang sering mengalami data hilang.

Tujuan

1. Menyelesaikan permasalahan Data Hilang
2. Menguji keefektifan pendekatan fungsi Closure
3. Mengimplementasikan fungsi Closure pada Python

Metode

1. Fungsi Closure Python

   Closure merupakan istilah pemrograman yang umum untuk mendeklarasikan suatu fungsi di dalam suatu fungsi(fungsi bersarang). Penerapan closure berguna ketika blok kode perlu dieksekusi beberapa kali tetapi hanya terjadi dalam fungsi tertentu atau eksekusi terjadi setelah pemanggilan fungsi tertentu.

Closure juga dapat digunakan untuk mengimplementasikan fungsi yang membutuhkan parameter yang berubah secara dinamis. Dalam pemrograman berbasis fungsi, closure dapat dibuat dengan menggunakan sintaks yang sederhana.

2. Teknik Data Imputer

Teknik Data Imputer adalah cara ataupun metode untuk 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 (Qin et al., n.d.,).

Dalam teknik data impurter ada beberapa pendekatan yang dapat dilakukan dengan yaitu pendekatan dengan regresi linier, pendekatan interpolasi dan pendekatan deskripsi statistik, dengan mempertimbangkan cara-cara untuk mengisi maupun memperkirakan nilai yang hilang dalam sebuah dataset. Terdapat juga pendekatan umum yang sering digunakan dalam data imputer seperti, pengisian nilai rata-rata, median dan modus.

a. Pendekatan Regresi Linear

Pendekatan regresi linier pada data imputer adalah proses untuk mengisi missing values dalam data dengan menggunakan model regresi linier. Model regresi linier ganda (MLR) adalah model yang digunakan untuk menangani missing values dalam data. Dalam analisis digunakan model MLR menggunakan chained equation (MICE) untuk menangani missing values. MICE adalah metode imputasi yang berantai sehingga meminimalisir timbulnya data yang tidak masuk akal.

Dalam pendekatan regresi linier pada data imputer, data imputasi dilakukan dengan menggunakan model regresi linier yang sesuai dengan tingkatan kompleksitas data yang akan dianalisis. Model regresi linier sederhana (OLS) dan regresi linier ganda (MLR) merupakan contoh dari model yang dapat digunakan untuk menangani missing values dalam data (Syakarna, 2014).

Penerapan regresi linier dalam pemrograman sering kali melibatkan penggunaan pustaka atau kerangka statistik seperti NumPy atau SciPy dalam bahasa pemrograman seperti Python. Dengan menggunakan library ini dapat dengan mudah melakukan regresi linier dan menerapkannya untuk memecahkan masalah terkait analisis atau prediksi berbasis data.

b. Pendekatan Interpolarisasi

Ketika kita berhadapan dengan data, sering kali data tersebut tidak tersaji dengan legkap atau terdapat data yang hilang *(missing value)*. Hal tersebut dapat terjadi karena kesalahan pada manusia sebagai pengamat data ataupun keterbatasan kemampuan alat ukur dalam mengolah data.

Selain itu, terdapat pula *outlier* atau nilai yang berbeda jauh dengan mayoritas data yang kita dapatkan. Nilai tersebut akan menentukan hasil analisis atau uji statistik yang akan kita lakukan. Namun, banyak cara untuk dapat menangani beberapa hal tersebut, salah satunya dengan melakukan interpolasi terhadap data yang kita miliki tersebut.

Interpolasi adalah proses menebak nilai data dengan memperhatikan data lain yang kita miliki. Pendekatan ini berguna untuk mencari nilai suatu variabel yang hilang pada rentang data yang diketahui. Metode ini menggunakan pendekatan berdasarkan kecenderungan dari sederet data atau nilai-nilai yang disajikan pada data.

Secara umum, saat digunakan dalam Python, interpolasi terbagi menjadi tiga jenis, yaitu interpolasi linear, interpolasi kuadratik, serta interpolasi berderajat tinggi (Rahmad, 2018).

c. Pendekatan Deskripsi Statistik

Ilmu Statistika dapat mempelajari cara penyajian data suatu penelitian dengan Deskripsi Statistik. Data yang ditampilkan akan lebih ringkas dan lebih deskriptif sehingga dapat memberikan informasi yang baik dan dapat dibaca dengan mudah (Walpole, 1995). Penyajian deskripsi statistik dapat ditampilkan melalui mean, median, serta modus.

Mean atau nama lainnya rata-rata dilambangkan dengan tanda x yang diberi garis di atasnya($\overline{x}$), yang kita sebuat dengan x bar, sedangkan untuk sampel dilambangkan dengan x (Kuswanto, 2012). Mean dapat dihitung dengan rumus sebagai berikut :

$$X = \frac{\Sigma x}{n}$$

Median merupakan nilai tengah dari pemusatan data yang membagi suatu data menjadi setengah data terkecil dan terbesarnya. Namun, median hanya dapat ditemukan jika data sudah terurut nilainya dari yang terkecil sampai terbesar. Median dapat dihitung dengan rumus sebagai berikut :

- Data tunggal ganjil

  $Me = \frac{X(n+1)}{2}$

- Data tunggal genap

  $Me = \frac{((X\frac{n}{2}) + (X\frac{n}{2}+1))}{2}$

Modus merupakan nilai yang paling sering muncul dalan suatu data statistika (Agus, 2007). Modus dapat digunakan untuk menentukan sampel dari suatu populasi dalam statistika. Perhitungan modus dapat diterapkan pada data numerik maupun data kategorik.

Pembahasan

1. Pseudocode

JavaScript

Function: data_imputer(strategy='mean')

Purpose: Fills missing values (NaNs) in a Pandas DataFrame or Series with specified imputation strategies.

Steps:

1. Define a nested function named `impute(data)`

```
def data_imputer(strategy='mean'):
    def impute(data):
        # Check data type
        if isinstance(data, pd.DataFrame):
            # Iterate through columns
            for col in data.columns:
                if pd.api.types.is_numeric_dtype(data[col]):
                 # Apply imputation based on strategy (mean, median, or mode)
                    if strategy == 'mean':
                        data[col].fillna(data[col].mean(), inplace=True)
                    elif strategy == 'median':
                        data[col].fillna(data[col].median(), inplace=True)
                    elif strategy == 'mode':
                        data[col].fillna(data[col].mode().iloc[0], inplace=True)
                    else:
                        raise ValueError("Unsupported strategy. Supported strategies are 'mean', 'median', and 'mode'.")
        elif isinstance(data, pd.Series):
            if pd.api.types.is_numeric_dtype(data):
                # Apply imputation based on strategy (mean, median, or mode)
                if strategy == 'mean':
                    data.fillna(data.mean(), inplace=True)
                elif strategy == 'median':
```

```
                data.fillna(data.median(), inplace=True)
            elif strategy == 'mode':
                data.fillna(data.mode().iloc[0], inplace=True)
            else:
                raise ValueError("Unsupported strategy. Supported strategies are 'mean', 'median', and 'mode'.")
        else:
            raise ValueError("Unsupported data type. Supported data types are numeric DataFrame and Series.")
        return data
    return impute
```

2. Kode Fungsi Data Imputer

Python

```
import pandas as pd
import numpy as np

def data_imputer(strategy='mean'):
    def impute(data):
        if isinstance(data, pd.DataFrame):
            for column in data.columns:
                if pd.api.types.is_numeric_dtype(data[column]):
                    if strategy == 'mean':
                        data[column] = data[column].fillna(data[column].mean())
                    elif strategy == 'median':
                        data[column] = data[column].fillna(data[column].median())
                    elif strategy == 'mode':
                        data[column] = data[column].fillna(data[column].mode()[0])
                    else:
                        raise ValueError("Unsupported imputation strategy")
        elif isinstance(data, pd.Series):
            if pd.api.types.is_numeric_dtype(data):
                if strategy == 'mean':
                    data = data.fillna(data.mean())
                elif strategy == 'median':
```

```
            data = data.fillna(data.median())
        elif strategy == 'mode':
            data = data.fillna(data.mode()[0])
        else:
            raise ValueError("Unsupported imputation strategy")
    else:
        raise ValueError("Unsupported data type. Expecting Pandas DataFrame or Series.")
    return data

return impute
```

3. Penerapannya pada Dataset Curah Hujan
   Dataset Curah Hujan Provinsi Lampung Tahun 2022
   Dataset yang digunakan dalam studi kasus kali ini berupa data jumlah curah hujan di provinsi lampung pada tahun 2022 yang dikumpulkan dari situs resmi Badan Pusat Statistik Provinsi Lampung.

Bulan	Jumlah Curah Hujan	Rata-rata Suhu Udara	Rata-rata Kelembaban Udara
Januari	Tinggi	34,2	100
Februari	Menengah	34,4	100
Maret	Menengah	34,6	100
April	Menengah	NuN	98
Mei	Menengah	35	100
Juni	Menengah	34,2	NaN
Juli	Rendah	33,4	98
Agustus	Menengah	34,2	98
September	Rendah	34	96
Oktober	Menengah	34,4	98
November	Menengah	35	99
Desember	Menengah	33,6	100

Menggunakan Mean

```python
impute_mean = data_imputer('mean')
impute_mean(df)
```

Dengan menggunakan fungsi data_imputer dengan strategy pendekatan nilai rata-rata (mean), didapatlah nilai yang melengkapi datasetnya sebagai berikut:

Bulan	Jumlah Curah Hujan	Rata-rata Suhu Udara	Rata-rata Kelembaban Udara
Januari	Tinggi	34,2	100
Februari	Menengah	34,4	100
Maret	Menengah	34,6	100
April	Menengah	34.272727	98
Mei	Menengah	35	100
Juni	Menengah	34,2	98.818182
Juli	Rendah	33,4	98
Agustus	Menengah	34,2	98
September	Rendah	34	96
Oktober	Menengah	34,4	98
November	Menengah	35	99
Desember	Menengah	33,6	100

Menggunakan Median

```python
impute_median = data_imputer('median')
impute_median(data)
```

Dengan menggunakan fungsi data_imputer dengan strategy pendekatan nilai rata-rata (mean), didapatlah nilai yang melengkapi datasetnya sebagai berikut:

Bulan	Jumlah Curah Hujan	Rata-rata Suhu Udara	Rata-rata Kelembaban Udara
Januari	Tinggi	34,2	100
Februari	Menengah	34,4	100
Maret	Menengah	34,6	100

April	Menengah	34,2	98
Mei	Menengah	35	100
Juni	Menengah	34,2	99
Juli	Rendah	33,4	98
Agustus	Menengah	34,2	98
September	Rendah	34	96
Oktober	Menengah	34,4	98
November	Menengah	35	99
Desember	Menengah	33,6	100

Menggunakan Modus

```python
impute_mode = data_imputer('mode')
impute_mode(data)
```

Dengan menggunakan fungsi data_imputer dengan strategy pendekatan nilai rata-rata (mean), didapatlah nilai yang melengkapi datasetnya sebagai berikut:

Bulan	Jumlah Curah Hujan	Rata-rata Suhu Udara	Rata-rata Kelembaban Udara
Januari	Tinggi	34,2	100
Februari	Menengah	34,4	100
Maret	Menengah	34,6	100
April	Menengah	34,2	98
Mei	Menengah	35	100
Juni	Menengah	34,2	100
Juli	Rendah	33,4	98
Agustus	Menengah	34,2	98
September	Rendah	34	96
Oktober	Menengah	34,4	98
November	Menengah	35	99
Desember	Menengah	33,6	100

Kesimpulan

Pada praktikum diperlukan data pada dataset,yaitu menggunakan teknik imputasi data. Kami mengusulkan penggunaan fungsi closure pada Python untuk melakukan imputasi data dengan metode statistik seperti mean, median, dan mode. Dengan fungsi closure, kita dapat membuat fungsi yang fleksibel dan dapat digunakan kembali dengan parameter yang berbeda sesuai kebutuhan. Melalui eksperimen pada dataset curah hujan Provinsi Lampung periode 2022, dapat disimpulkan bahwa pendekatan ini berhasil mengatasi kekurangan data dengan tingkat akurasi yang memuaskan. Penggunaan fungsi closure memungkinkan implementasi berbagai metode imputasi data dengan mudah, hanya dengan satu fungsi yang dapat disesuaikan. Penerapan teknik ini pada dataset curah hujan Provinsi Lampung menunjukkan bahwa dengan menggunakan nilai rata-rata (mean), median, dan mode, kita dapat mengisi kekurangan data dengan nilai yang sesuai. Hal ini menunjukkan bahwa metode data imputer dengan fungsi closure dapat diterapkan pada berbagai jenis dataset, bukan hanya pada data cuaca Provinsi Lampung. Dengan demikian, pendekatan ini memiliki potensi untuk meningkatkan kualitas dan kegunaan berbagai dataset yang terpengaruh oleh kekurangan data, dan bisa menjadi solusi yang efektif dalam pengolahan data.

Referensi

Agus, Nuniek Avianti. 2007. Mudah Belajar Matematika. Jakarta: Pusat Perbukuan Departermen Pendidikan Nasional. hlm. 138. ISBN 9794628182.

Kuswanto, D. 2012. Statistik Untuk Pemula dan Orang Awam. Jakarta:Penerbit Laskar Aksara.

Qin, Y., Zhu, X., Zhang, J., & Zhang, C. (n.d.). *Semiparametric optimization for missing data imputation*. *Appl. Intell.,*, *27*(no. 1), 79–88. 5765-13494-1-SM

Rahmad, C., dkk. 2018. Metode Numerik. UPT Percetakan dan Penerbitan Polinema. hal 68-70.

Syakarna, N. F. R. 2014. Analisis Metode Regresi Untuk Imputasi Data Pada Survei Sampel. Data Imputasi. http://etheses.uin-malang.ac.id/7046/1/09610045.pdf

Walpole, Ronald E., Raymond H Myers.; 1995. Ilmu Peluang Dan Statistika untuk Insinyur dan Ilmuawan, edisi ke-4, Penerbit ITB, Bandung.