Implementasi Matriks Korelasi pada Dataset Diabetes

Arafi Ramadhan Maulana (122450002)[1], Dwi Ratna Anggraeni (122450008)[2], Raid Muhammad Naufal (122450027)[3], Rayan Koemi Karuby (122450038)[4], Muhammad Deriansyah Okutra (122450101)[5]

Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera

Email :
arafi.122450002@student.itera.ac.id [1] , *dwi.122450008@student.itera.ac.id* [2] , *raid.122450027@student.itera.ac.id* [3] , *rayan.122450038@student.itera.ac.id* [4] , *mderiansyah.122450101@student.itera.ac.id* [5]

1. Pendahuluan

Matriks korelasi merupakan matriks pasti semi positif yang dapat menggambarkan ketergantungan yang beragam antar kumpulan data. Dalam kasus dengan multi variabel atau kasus dengan banyak variabel sekunder yang sulit untuk dilakukan prediksi distribusi spasial dan ketergantungan antar variabelnya, matriks korelasi menjadi solusi untuk menggambarkan ketergantungannya. (Siri & Deutsch, 2012)

Matriks korelasi dapat digunakan untuk meringkas data yang besar serta mengidentifikasi pola dan mengambil keputusan berdasarkan data tersebut. Dengan matriks korelasi dapat dilihat variabel mana yang lebih berkorelasi dengan variabel tertentu dan hasilnya dapat divisualisasikan. (Borate, 2017)

Prinsip dari matriks korelasi adalah mencari hubungan antar variabel mana yang mempunyai korelasi paling erat, lalu salah satu atribut dari yang saling berkorelasi tersebut dihilangkan. Setiap sel pada matriks korelasi berisi koefisien korelasi.Matriks korelasi dapat digabungkan dengan jenis analisis statistik lain. Pada praktikum kali ini digunakan matriks korelasi dengan peta *Heatmap* yang dapat menggambarkan hubungan korelasi tiap variabel, sehingga didapatkan variabel yang relevan terhadap variabel output (Amiruddin & Ishak, 2022)

Teknik Correlation Matrix with *Heatmap* diimplementasikan pada dataset diabetes yang bersumber dari National Institute of Diabetes and Digestive and Kidney Diseases. Objektif dari dataset tersebut adalah untuk memprediksi secara diagnosa apakah pasien mengisap diabetes berdasarkan beberapa pengukuran diagnostik dalam dataset tersebut. Dalam dataset diabetes terdiri dari beberapa variabel prediktor medis dan satu variabel target, yaitu *Outcome*. Variabel prediktor terdiri dari angka kehamilan, BMI, tingkat insulin, umur, dan lainnya (UCI Machine Learning Kaggle Team, 2016)

2. Metode

2.1. Seaborn

Seaborn adalah salah satu *library* dalam *python* yang digunakan untuk visualisasi data dengan membuat grafik dan statistik, *library* ini dibangun berdasarkan *library matplotlib.*

2.1.1. heatmap()

Heatmap adalah teknik visualisasi dua dimensi yang menunjukkan variasi dalam besarnya fenomena tertentu dalam hal warna yang diwakili warna berbeda, ini memudahkan untuk mengidentifikasi fitur mana yang paling terkait dengan variabel terkait.

2.2. Matplotlib

Matplotlib adalah sebuah *library* yang komprehensif yang berfungsi memvisualisasikan statis, animasi, dan interaktif dalam *python*. Ada beberapa macam grafik yang dapat dibuat oleh *matplotlib* misalnya grafik lingkaran, garis, batang, histogram, dan lain-lain.

2.2.1. figure()

Figure adalah area kerja atau ruang kosong dalam objek visual yang digunakan untuk membuat plot, misalnya dalam menggambar di kertas, maka kertas itulah yang disebut *figure*. Untuk tiap *figure* mempunyai ukuran, judul, dan atribut lainnya.

2.2.2. show()

Show adalah salah satu fitur dalam pustaka *matplotlib*, dimana setelah plot selesai dibuat maka untuk menampilkan plot tersebut secara interaktif dapat menggunakan fungsi *show()*.

2.3. Pandas

Pandas adalah salah satu *library* yang ada dalam *python* yang berlisensi BSD dan open source, didalamnya terdapat struktur data dan analisis data yang mudah untuk digunakan. Berikut beberapa fungsi dari *pandas* yaitu menyelaraskan data sebelum dibandingkan ataupun penggabungan *dataset,* menangani data hilang, dan lain-lain

2.3.1. DataFrame()

Dataframe adalah struktur data *pandas* yang terdiri dari beberapa kolom yang berurutan dengan nama dan jenis. dan memudahkan dalam membaca sebuah file dan menjadikan sebuah *tabel.* Beberapa operasi yang digunakan distinct, join, group by dan lain sebagainya. Operasi tersebut digunakan untuk mengelola data.

2.3.2. corr()

Metode corr() adalah metode dalam pustaka pandas yang digunakan untuk menghitung hubungan antar setiap kolom dalam kumpulan *dataset,* outputnya berupa angka yang mempresentasikan seberapa baik hubungan antar kolom.

2.3.3. read_csv()

Read_csv adalah salah satu fungsi pada *pandas* yang digunakan untuk membaca file csv sehingga kita dapat memanipulasi data dari file csv tersebut.

2.3.4. replace()

Fungsi tersebut digunakan untuk mengelola data berbentuk *string,* dimana dapat menggantikan sebagian karakter atau keseluruhan *string* menjadi bentuk yang kita inginkan.

2.4. correlation_matrix_generator()

Generator matriks korelasi adalah metode yang mudah dan cepat untuk mengidentifikasi dan menganalisis hubungan antara dua variabel. Matriks

korelasi merupakan matriks bujur sangkar yang memuat korelasi koefisien antara semua kombinasi variabel. Dengan beberapa klik, generator akan membuat visualisasi berkode warna dari data.

2.5. correlation_plot()

Plot korelasi adalah visualisasi dalam bentuk grafik yang digunakan untuk menggambarkan hubungan linear antara dua variabel. Pada grafik, nilai suatu variabel ditunjukkan pada sumbu x, sedangkan nilai variabel lain ditunjukkan pada sumbu y, ini berfungsi menunjukkan hubungan positif, negatif, bahkan tidak ada hubungan.

3. Pembahasan

Pada studi kasus ini data yang digunakan merupakan dataset *Pima Indians Diabetes* yang memiliki variabel (*Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, BMI, *DiabetesPedigreeFunction*, *Age*, *Outcome*). Variabel *pregnancies* mempresentasikan berapa kali wanita tersebut hamil selama hidupnya, variabel *glucose* mempresentasikan konsentrasi plasma glukosa, variabel *blood pressure* mempresentasikan tekanan darah diastolik dalam (mmHg), variabel *skin thickness* mempresentasikan perkiraan lemak tubuh, variabel *insulin* mempresentasikan tingkat *insulin* serum, variabel BMI mempresentasikan indeks massa tubuh, variabel *diabetes pedigree function* mempresentasikan indikator riwayat diabetes dalam keluarga, variabel *age* mempresentasikan umur dalam tahun dan variabel *outcome* mempresentasikan kelas berupa *case* yang memiliki arti penderita diabetes dan *control* yang memiliki arti bukan penderita diabetes**.**

```
[1]  import seaborn as sns
     import matplotlib.pyplot as plt
     import pandas as pd
```

Gambar 1. *Import library*

Pada aplikasi ini *library-library* yang digunakan adalah *library seaborn, matplotlib,* dan *pandas. Seaborn* digunakan untuk memvisualisasi data yang dibangun dengan *library matplotlib* yang digunakan untuk memvisualisasi grafik data. *Pandas* digunakan untuk memanipulasi dan analisis data. Pada aplikasi ini *library seaborn* dan *matplotlib* digunakan untuk membuat atau memvisualisasikan plot matriks korelasi dan *library pandas* digunakan untuk meng-*import* data serta memanipulasi data.

```
[2]  def correlation_matrix_generator():
       def correlation_plot(data):
         df = pd.DataFrame(data)
         corr_matrix = df.corr()
         plt.figure(figsize=(10, 10))
         sns.heatmap(corr_matrix, annot=True)
         plt.show()

       return correlation_plot
```

Gambar 2. Fungsi *Correlation Matrix Generator*

Selanjutnya buatlah fungsi *closure* pada python, pada kode dibawah ini biasanya digunakan untuk menghitung korelasi antara variabel dalam data dan memplotnya dalam bentuk *heatmap*. Fungsi ini mengambil data sebagai argumen, menghitung korelasi dengan fungsi *corr()*, dapat diperhatikan pada fungsi *correlation_matrix_generator* hanya mengembalikan fungsi *correlation_plot*, jadi perlu memanggilnya dengan argumen data, sehingga akan menjalankan fungsi *correlation_plot.* Pada fungsi *correlation_plot* data

didefinisikan ulang dengan fungsi *DataFrame(data)* yang akan merubahnya ke dalam bentuk *data frame* sebagai *df,* lalu mendefinisikan *corr_matrix* sebagai fungsi *corr()* yang akan menghasilkan nilai koefisien korelasi antar variabel-variabel pada *df,* kemudian membuat plot menggunakan fungsi *figure(figsize=(10, 10))* dan fungsi *heatmap(corr_matrix, annot=True),* dan menampilkan data menggunakan fungsi *show().*

```
[3]  data = pd.read_csv("diabetes.csv")
```

Gambar 3. *Import* Dataset

Setelah membuat fungsi *correlation matrix generator*, langkah selanjutnya adalah *import* dataset *Pima Indians Diabetes* yang didefinisikan sebagai data dengan menggunakan fungsi *read_csv("diabetes.csv").*

```
[4]  data['Outcome'] = data['Outcome'].replace({'Case': 1, 'Control': 0})
```

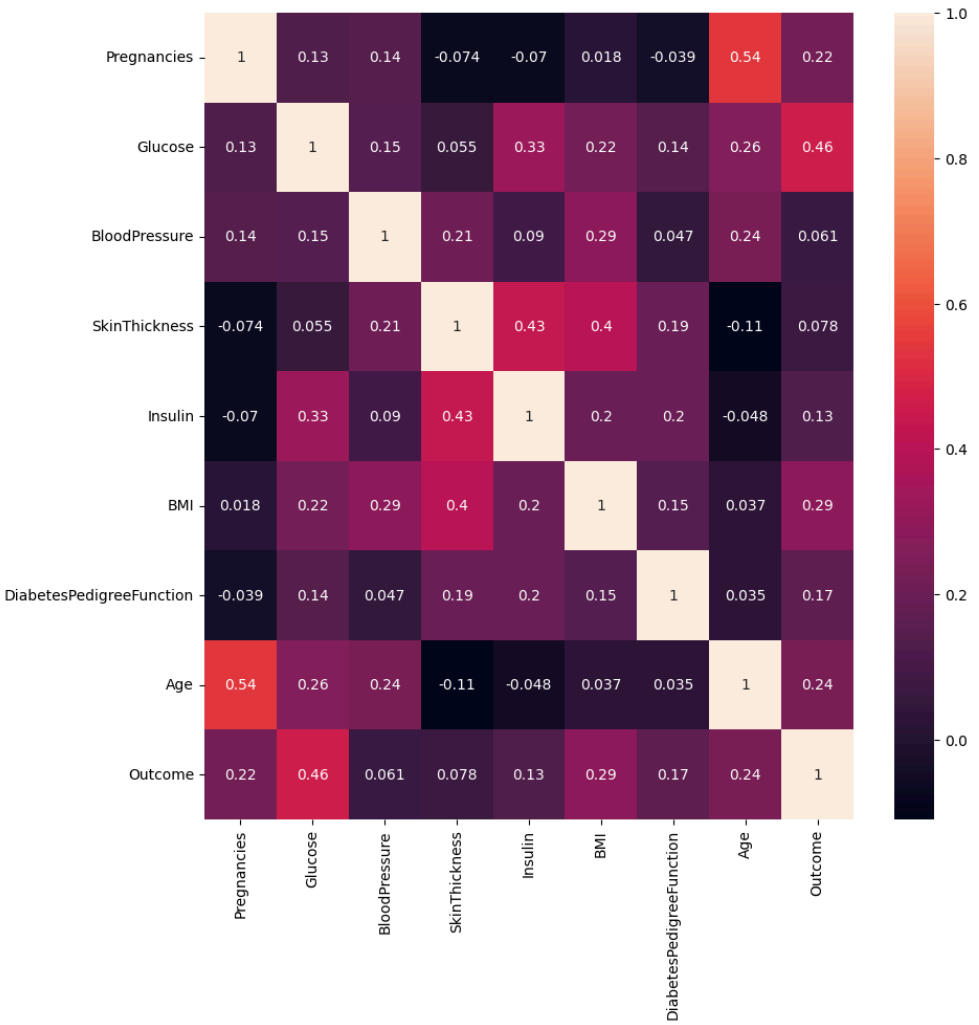
Gambar 4. Fungsi untuk Mengubah Nilai DataFrame

Saat dataset berhasil di-*import,* langkah selanjutnya adalah merubah nilai-nilai data pada kolom atau variabel *outcome* menjadi angka dengan *case* menjadi satu dan *control* menjadi nol dengan menggunakan fungsi *replace({'Case': 1, 'Control': 0})*, perubahan ini digunakan untuk analisis lebih lanjut dan mempermudahnya dalam membuat matriks korelasi.

```
[5]  correlation = correlation_matrix_generator()
     correlation(data)
```

Gambar 5. Korelasi antar Variabel

Setelah nilai-nilai pada kolom *outcome* diubah, maka fungsi *correlation_matrix_generator* dapat didefinisikan sebagai *correlation* untuk mempermudah penggunaan fungsi *correlation matrix generator.* Pada fungsi ini bertujuan untuk mencari korelasi antar variabel di dalam dataset *Pima Indians Diabetes* yang nantinya akan menghasilkan output berupa *plot heatmap* matriks korelasi dari data.



Gambar 6. *Plot Heatmap* Matriks Korelasi

Berdasarkan Gambar 6 yang menunjukkan untuk setiap variabel yang memiliki nilai koefisien korelasi mendekati positif satu dan negatif satu, maka hubungannya akan semakin kuat, begitu juga sebaliknya, apabila nilai koefisien korelasinya mendekati nol maka hubungannya semakin lemah. Pada hubungan variabel-variabel yang memiliki nilai koefisien korelasi di atas 0.40 atau di bawah -0.40 adalah variabel *age* dan *pregnancies* dengan nilai koefisien korelasi sebesar 0.54, variabel *glucose* dan *outcome* dengan nilai koefisien korelasi sebesar 0.46, variabel *insulin* dan *skin thickness* dengan nilai koefisien korelasi sebesar 0.43, serta variabel BMI dan *skin thickness* dengan nilai koefisien korelasi sebesar 0.4 yang menyatakan bahwa variabel-variabel berikut memiliki hubungan antar variabelnya. Sedangkan untuk hubungan variabel-variabel lainnya memiliki nilai koefisien korelasi mendekati nol yang menyatakan hubungannya lemah.

4. Kesimpulan

Generator matriks korelasi dapat digunakan untuk membantu pengguna mengidentifikasi hubungan dan kuatnya antara variabel-variabel, sehingga kita dapat memahami data yang kita analisis lebih baik.

Pada matriks korelasi yang telah dilakukan menggunakan dataset *Pima Indians Diabetes* yang memiliki sembilan variabel, diketahui hubungan antara variabel-variabel tersebut dengan menggunakan fungsi *correlation matrix generator*, berdasarkan hasil visualisasi pada plot heatmap dapat dilihat variabel-variabel ditunjukkan pada sumbu x dan sumbu y, pada plot ini nilai positif menunjukkan hubungan searah antara variabel, sedangkan

nilai negatif menunjukkan hubungan berlawan arah antar variabel. Nilai koefisien korelasi menunjukkan kuatnya hubungan antara variabel-variabel, hubungan kuat antara variabel pada saat nilai koefisien korelasi mendekati nilai positif satu atau negatif satu, sedangkan hubungan lemah antara variabel pada saat nilai koefisien korelasi mendekati nilai nol. Pada matriks korelasi variabel-variabel memiliki nilai koefisien korelasi sebesar 0.54 untuk variabel age dengan pregnancies yang menyatakan hubungan antar variabel bersifat kuat dan nilai koefisien korelasi variabel lain berada dibawah 0.50 atau mendekati nol yang menyatakan hubungan antar variabel bersifat lemah.

5. Daftar Pustaka

Amiruddin, & Ishak, R. (2022, Juli). Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Correlation Matrix With Heatmap. *Jambura Journal of Electrical and Electronics Engineering*, *4*(2), 169. https://ejurnal.ung.ac.id/. -

Borate, R. (2017). *Correlation Matrix*. -. -

Siri, M. H., & Deutsch, C. V. (2012). Some Thoughts on Understanding Correlation Matrices. In *CGG Annual Report* (14th ed., p. 408). Centre for Computational Geostatistics Alberta. -

UCI Machine Learning Kaggle Team. (2016, - -). *Pima Indians Diabetes Database*. Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database