Project Akhir Visual Analytics for Causal Modeling Analysis Pada Faktor Risiko Stroke

1. Ringkasan

Causal modeling analysis pada faktor risiko stroke memiliki tujuan untuk menganalisis faktor risiko utama yang mempunyai pengaruh terhadap kemungkinan seseorang terkena stroke dengan penggunaan dataset yang berasal dari dataset publik Kaggle yaitu *Stroke Prediction Dataset*. Dataset ini berisi 5.109 data pasien dengan 16 fitur. Proses analisis dimulai dengan eksplorasi data untuk mengidentifikasi korelasi antara variabel. Variabel seperti umur (*age*), riwayat penyakit jantung (*heart_disease*), riwayat hipertensi (*hypertension*), dan rata-rata kadar glukosa dalam darah (*avg_glucose_level*) memiliki hubungan yang paling signifikan dengan kejadian stroke. Pendekatan seperti *Causal Directed Acyclic Graphs (CDAG)* dan Bayesian Network digunakan untuk mengungkap hubungan kausal antar variabelnya.

Hasil analisis kausal menunjukkan bahwa faktor yang paling signifikan dalam perubahan probabilitas terjadinya stroke adalah riwayat penyakit jantung (*heart_disease*) dan hipertensi (*hypertension*). Metode intervensi dilakukan untuk mengukur dampak perubahan pada variabel tertentu, seperti usia lanjut, riwayat hipertensi, dan riwayat penyakit jantung, terhadap probabilitas stroke, hingga ditemukan variabel parents yang memiliki hubungan kausal (sebab-akibat) dengan stroke dengan peningkatan probabilitas sebesar 11% untuk variabel umur dan 10% untuk variabel riwayat penyakit jantung.

2. Pendahuluan

Meningkatnya kasus stroke yang juga sebagai salah satu penyebab utama kematian dan kecacatan di seluruh dunia berdasarkan data WHO, menjadikan salah satu trigger akan pentingnya kesadaran terhadap faktor risiko terjadinya stroke pada seseorang. Dengan melakukan eksplorasi hubungan kausal antara faktor faktor yang berkorelasi dengan kejadian stroke pada pasien yang tercatat, mendukung upaya pencegahan terjadinya stroke. Eksperimen ini memanfaatkan dataset publik yang tersedia pada platform Kaggle yang akan digunakan untuk memahami hubungan kausal antara variabel independen spesifik dengan variabel dependen stroke. Visualisasi yang dihasilkan diharapkan dapat memberikan insight mendalam terkait faktor risiko paling penting atau paling mempengaruhi akan terjadinya stroke.

3. Studi Kasus

Dataset yang digunakan dari dataset asli Kaggle Stroke Prediction Dataset miliki Fedesoriano dengan deskripsi variabel pada Tabel 1. Yang kemudian dilakukan preprocessing oleh Daniel Dobrenz hingga mendapatkan dataset bersih (cleaned) yang dapat dilihat pada Gambar 1. Dataset ini terdiri dari total data berjumlah 5109 data dengan banyaknya fitur/variabel sejumlah 16 fitur. Namun dataset ini merupakan dataset imbalance karena

terdapat 4860 data pasien yang merupakan bukan pasien stroke, dan 249 data pasien yang merupakan pasien stroke.

Variabel	Deskripsi
ID	Identifier unik
Gender	Jenis kelamin pasien (Male/Female)
Age	Umur pasien dalam tahun
Hypertension	Riwayat hipertensi (0: tidak ada, 1: ada)
Heart_disease	Riwayat penyakit jantung (0: tidak ada, 1: ada)
Ever_married	Status pernikahan (Yes/No)
Work_type	Jenis pekerjaan pasien ("children", "Govt_jov", "Never_worked", "Private", "Self-employed")
Residence_type	Tipe tempat tinggal (Urban/Rural)
Avg_glucose_level	Rata-rata kadar glukosa dalam darah pasien
BMI	Indeks massa tubuh pasien
Smoking_status	Kebiasaan merokok (formerly smoked, never smoked, smokes)
Stroke	Apakah pasien pernah mengalami stroke (0: tidak, 1: ya)

Tabel 1. Deskripsi Tabel Dataset Asli

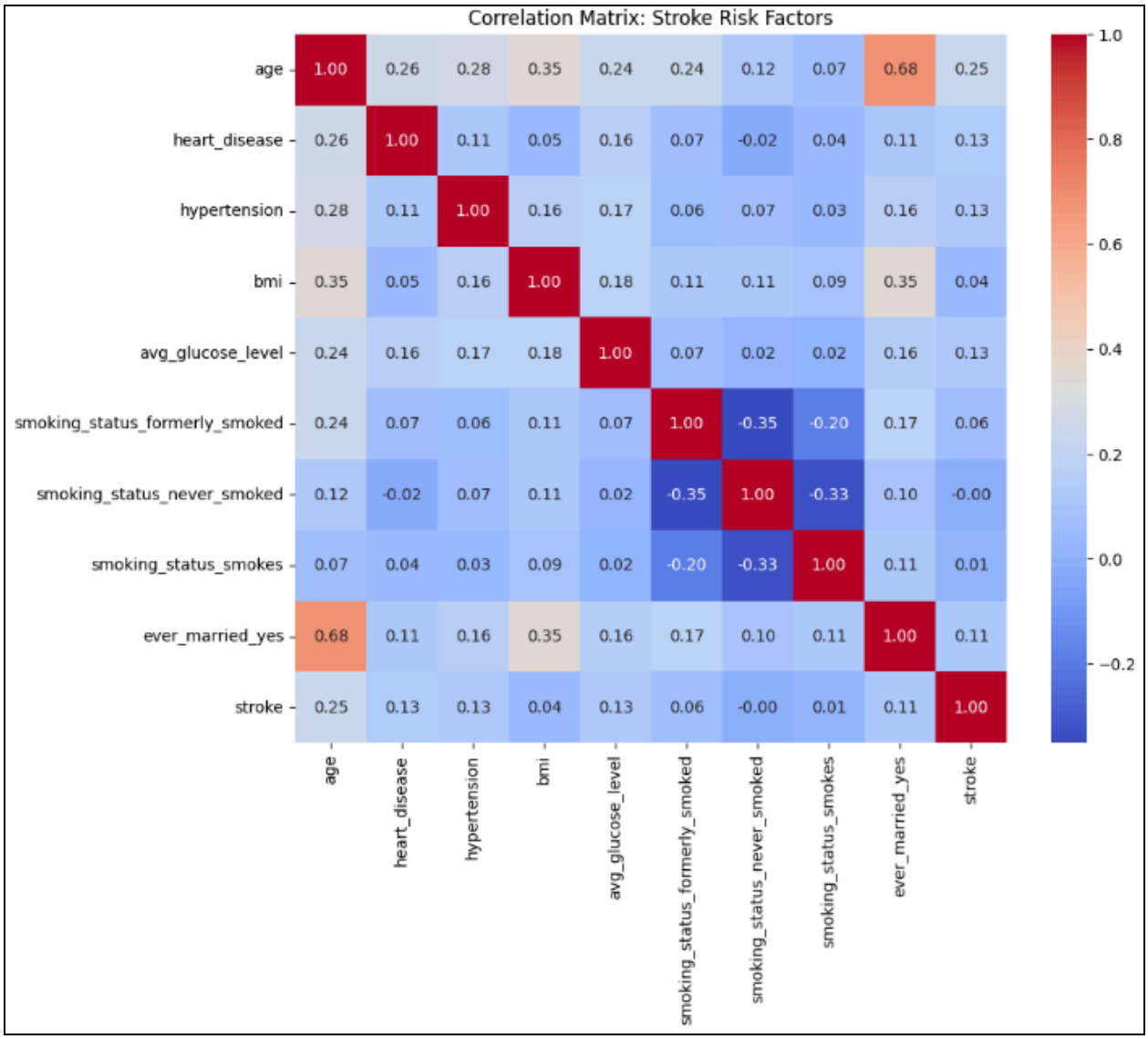
```
1. Info Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5109 entries, 0 to 5108
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   age                             5109 non-null   int64
 1   hypertension                    5109 non-null   int64
 2   heart_disease                   5109 non-null   int64
 3   avg_glucose_level               5109 non-null   int64
 4   bmi                             5109 non-null   int64
 5   stroke                          5109 non-null   int64
 6   gender_male                     5109 non-null   int64
 7   ever_married_yes                5109 non-null   int64
 8   work_type_never_worked          5109 non-null   int64
 9   work_type_private               5109 non-null   int64
 10  work_type_self_employed         5109 non-null   int64
 11  work_type_children              5109 non-null   int64
 12  residence_type_urban            5109 non-null   int64
 13  smoking_status_formerly_smoked  5109 non-null   int64
 14  smoking_status_never_smoked     5109 non-null   int64
 15  smoking_status_smokes           5109 non-null   int64
dtypes: int64(16)
memory usage: 638.8 KB
None
```

Gambar 1. Deskripsi Dataset Preprocessing

4. Analisis Data

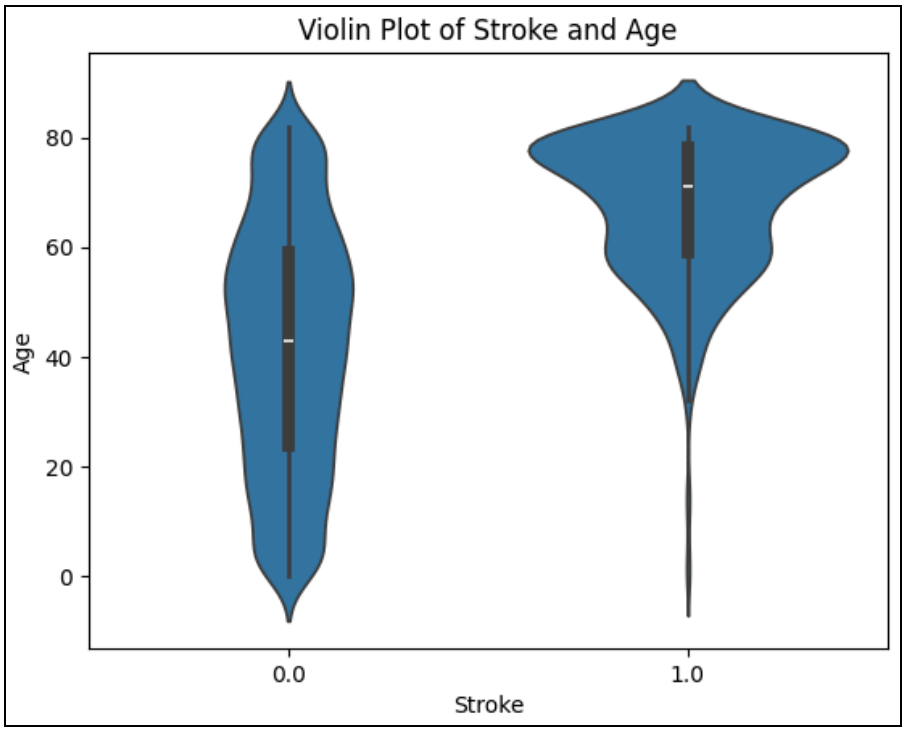
a. Eksplorasi data

Dataset stroke prediction ini telah mengalami preprocessing meskipun merupakan dataset yang imbalance namun akan tetap digunakan tanpa melakukan proses balancing data dengan menggunakan SMOTE demi menjaga keaslian dataset, sehingga dataset ini bisa langsung digunakan untuk proses analisis hubungan kausal. Pertama, kita perlu melihat korelasi antar variabel. Dari Gambar 2, yang merupakan heatmap korelasi antar variabel ini menunjukan bahwa variabel independen yang memiliki nilai korelasi paling tinggi dengan variabel dependen stroke adalah variabel age dengan total 0.25, nilai korelasi tinggi selanjutnya adalah heart_disease, hypertension, avg_glucose_level dengan total nilai korelasi sama yaitu 0.13. Kemudian ada pula variabel ever_married_yes yang memiliki nilai korelasi 0.11.

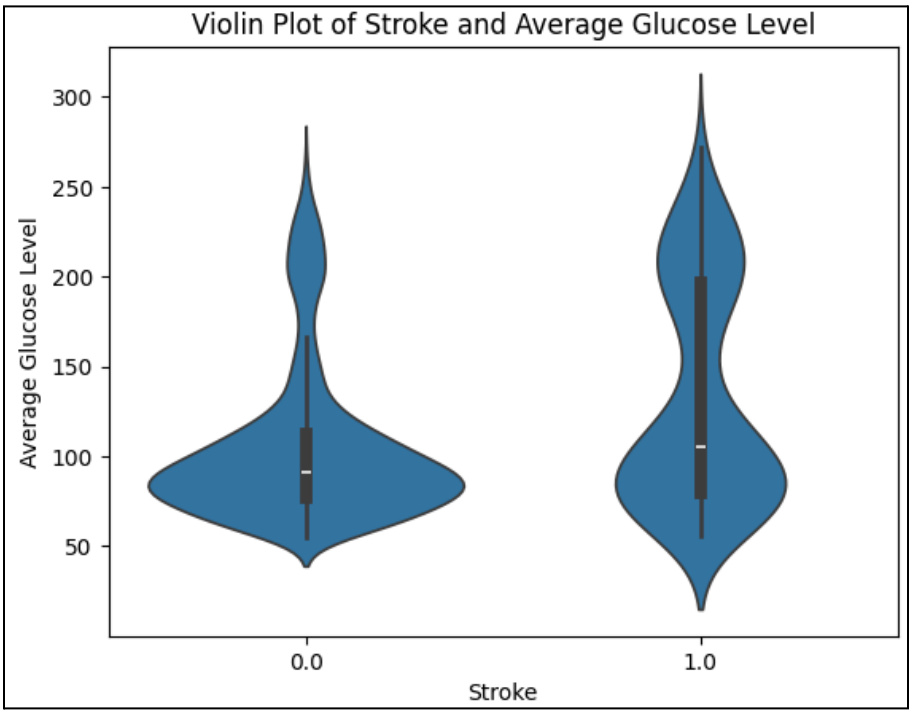


Gambar 2. Correlation Matrix antar variabel

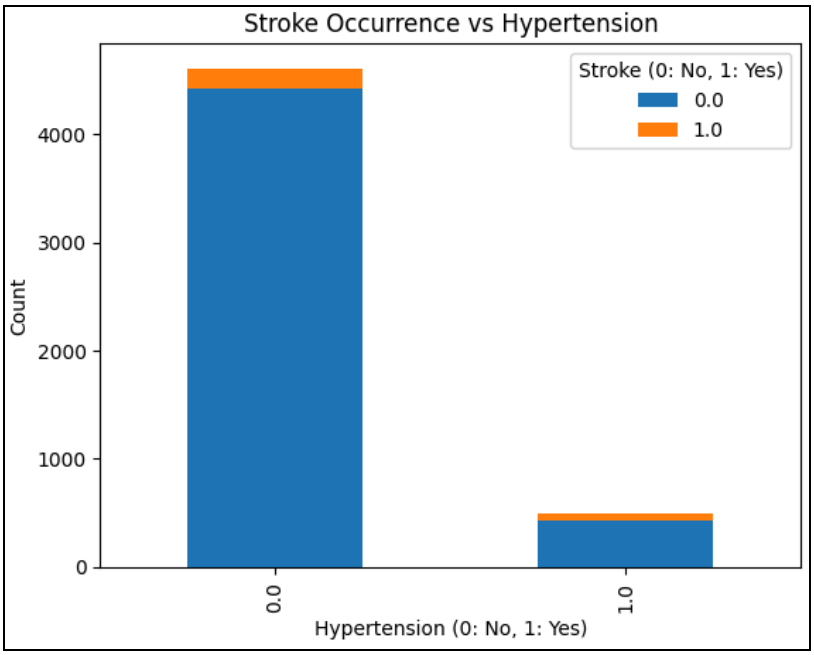
Gambar 3, 4, 5, dan 6 merupakan plot distribusi masing masing variabel dengan nilai korelasi tinggi (lebih dari 0.1) akan terjadinya stroke.



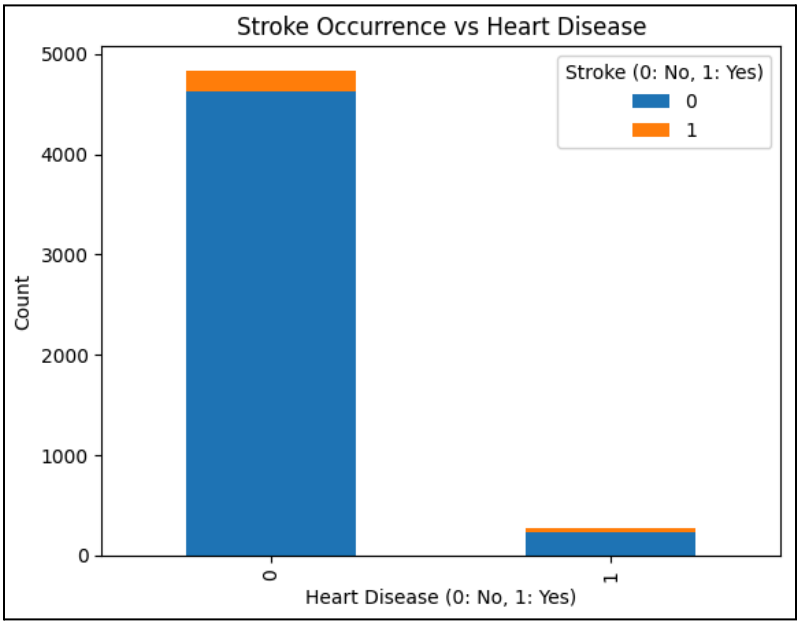
Gambar 3. Violin Plot Stroke Vs Age



Gambar 4. Violin Plot Stroke Vs Avg_Glucose_Level



Gambar 5. Stacked Bar Chart Stroke Vs Hypertension



Gambar 6. Stacked Bar Chart Stroke Vs Heart_Disease

b. Hitung Probabilitas Terjadinya Stroke

Tahap selanjutnya adalah menghitung nilai probabilitas akan terjadinya stroke dari seluruh dataset, yang mana didapatkan nilai sebesar 0.0458. Artinya kemungkinan terjadi stroke sebesar 4,58%. Dari korelasi tinggi pada point a, dihitung berapa nilai probabilitas terjadinya stroke menggunakan metode Teorema Bayes untuk masing masing variabel.

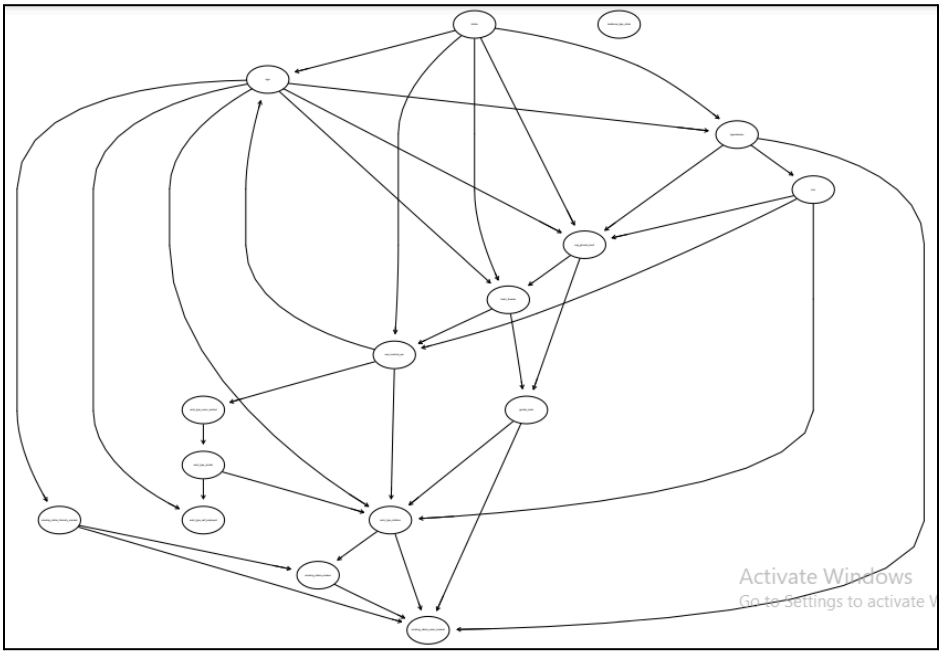
Didapatkan Tabel 2, dengan nilai probabilitas terjadinya stroke tertinggi adalah variabel heart_disease dengan nilai probabilitas 17% dan variabel hypertension sebesar 13%.

Variabel	Probability (%)
Age	7,637
Avg_Glucose_Level	6,091
Hypertension	13,253
Heart_Disease	17,028
Ever_Married_Yes	6,561

Tabel 2. Probabilitas antar Variabel akan terjadinya stroke

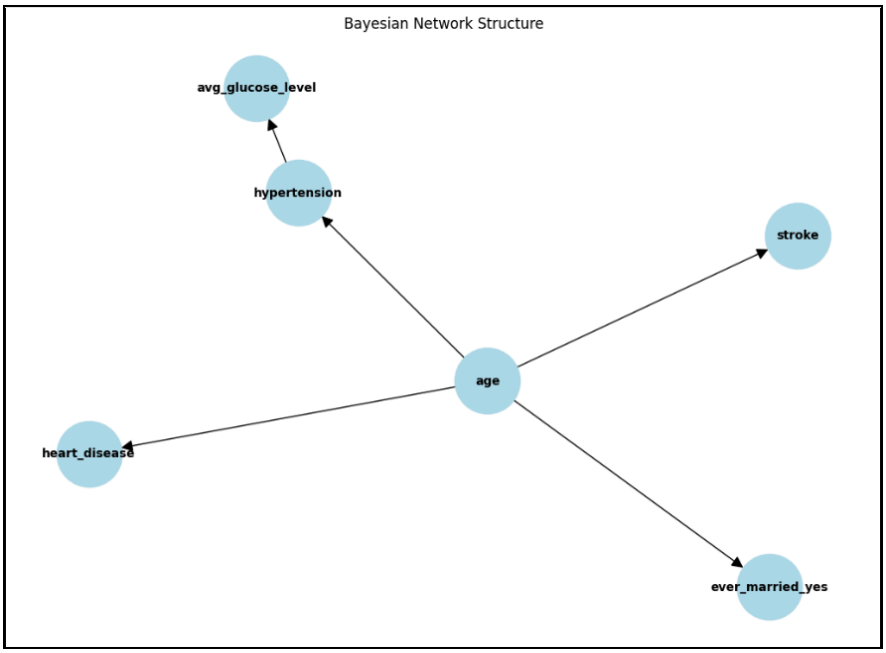
c. Causal Graph Diagram

Untuk melihat hubungan kausal sebenarnya, diperlukan Causal Directed Acyclic Graph (CDAG). Dengan menggunakan metode PC algorithm, didapatkan hasil CDAG pada Gambar 7 (guna kemudahan pembacaan gambar karena akan sulit dibaca walaupun telah diperbesar, penulis sertakan link github pada bagian rujukan). Pada gambar CDAG ditemukan bahwa stroke memiliki hubungan kausal yang ditunjukan dari adanya panah ke variabel-variabel age, avg_glucose_level, hypertension, heart_disease, ever_married_yes. Maka kelima variabel tersebutnya yang dinyatakan variabel yang mungkin memiliki pengaruh akan terjadinya stroke.



Gambar 7. Causal Directed Acyclic Graph

Dari hasil CDAG, tahap selanjutnya adalah pembuatan CDAG spesifik untuk variabel yang dipilih sebelumnya. Model Bayesian Network digunakan untuk penentuan CDAG variabel terpilih sehingga didapatkan Gambar 8. Dilihat pada gambar bagaimana hubungan kausalitas sebenarnya terjadi, dimana variabel age memiliki hubungan kausalitas langsung dengan variabel stroke, variabel seperti heart_disease, ever_married_yes, dan hypertension memiliki hubungan kausal tidak langsung dengan stroke dengan melalu variabel age. sedangkan variabel avg_glucose_level memiliki hubungan kausal jauh dengan melalui variabel hypertension dan variabel age.



Gambar 8. CDAG Variabel Terpilih

d. Estimasi causal effect

Estimasi effect menggunakan linear regression dari tiap variabel untuk variabel stroke ditunjukan pada Tabel 3 berikut:

Variabel	Estimasi Effek
Age	0.002527479143082108
Avg_Glucose_Level	0.00030903562254723516
Hypertension	0.038266849131546777
Heart_Disease	0.056993577743163035

Ever_Married_Yes	-0.04430669193244334

Tabel 3. Estimasi Efek Kausal tiap Variabel terhadap Variabel Stroke

Estimasi efek kausal menunjukan besarnya pengaruh kausal (mengalami peningkatan atau pengurangan) suatu variabel independen pada variabel dependen. Dalam hal ini, variabel dengan pengaruh signifikan adalah variabel heart_disease dengan peningkatan probabilitas sebesar 5,69% dan variabel hypertension sebesar 3,82%. Untuk variabel age dan avg_glucose_level memiliki pengaruh peningkatan probabilitas yang tidak signifikan, sedangkan variabel ever_married_yes memiliki pengaruh penurunan probabilitas terhadap variabel stroke namun masih tidak signifikan.

e. Intervensi

Intervensi adalah tindakan langsung untuk memodifikasi nilai variabel independen tertentu yang kemudian dianalisis bagaimana intervensi ini berpengaruh terhadap variabel dependen. Pada Gambar 9, dilihat bahwa dilakukannya percobaan intervensi untuk variabel age untuk kategori 2 (kategori untuk usia lanjut), probabilitas seseorang terkena stroke mengalami peningkatan sebesar 11,6%. Demikian pula percobaan untuk variabel hypertension atau heart_disease, untuk variabel hypertension dengan intervensi kategori pasien memiliki riwayat hipertensi maka mengalami peningkatan sebesar 8,6% dan untuk variabel heart_disease mengalami peningkatan untuk intervensi ketika pasien memiliki riwayat penyakit jantung sebesar 10,1%. Perubahan probabilitas sesudah dan sebelum intervensi ini dapat terlihat pada grafik bar chart pada Gambar 12.

```
Intervention: P(stroke | age=2):
+----------+---------------+
| stroke   |   phi(stroke) |
+==========+===============+
| stroke(0) |        0.8832 |
+----------+---------------+
| stroke(1) |        0.1168 |
+----------+---------------+

Intervention: P(stroke | hypertension=1):
+----------+---------------+
| stroke   |   phi(stroke) |
+==========+===============+
| stroke(0) |        0.9140 |
+----------+---------------+
| stroke(1) |        0.0860 |
+----------+---------------+

Intervention: P(stroke | heart_disease=1):
+----------+---------------+
| stroke   |   phi(stroke) |
+==========+===============+
| stroke(0) |        0.8984 |
+----------+---------------+
| stroke(1) |        0.1016 |
+----------+---------------+
```

Gambar 9. Probabilitas Hasil Intervensi per Variabel

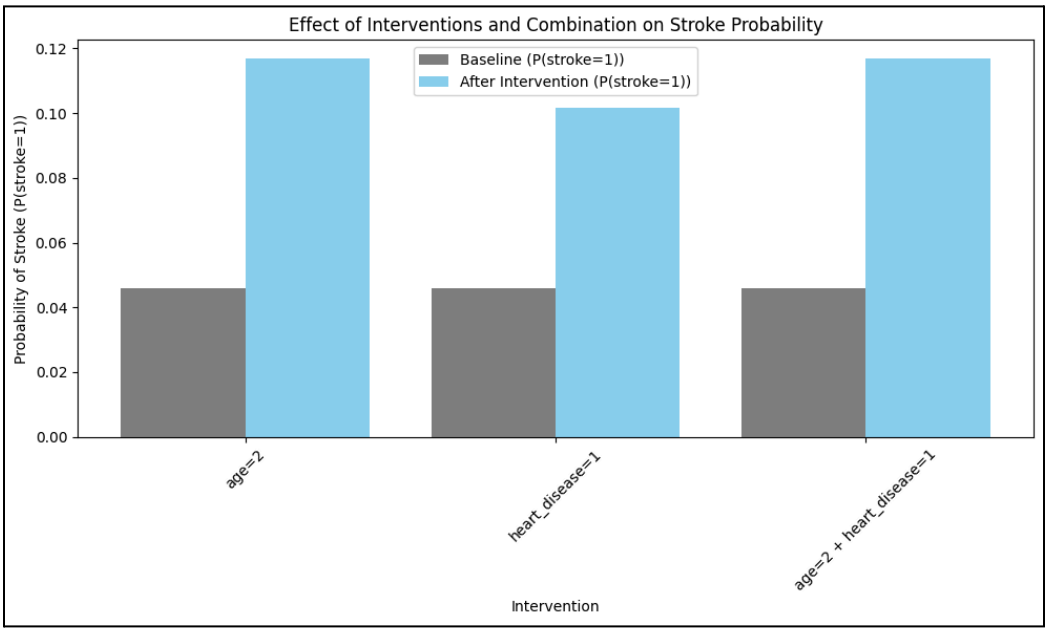
Dilakukan percobaan lain dengan mengkombinasikan 2 variabel yaitu age dan heart_disease. Ditemukan peningkatan probabilitas setelah intervensi sebesar 11,68% jika variabel age ada pada kategori 2 dan juga memiliki riwayat jantung. Peningkatan dan bagaimana perbandingan perubahan sebelum dan sesudah intervensi untuk 2 variabel ini dapat dilihat pada Gambar 10 dan Gambar 12.

```
Intervention: P(stroke | age=2, heart_disease=1):
+----------+---------------+
| stroke   |   phi(stroke) |
+==========+===============+
| stroke(0) |        0.8832 |
+----------+---------------+
| stroke(1) |        0.1168 |
+----------+---------------+
```

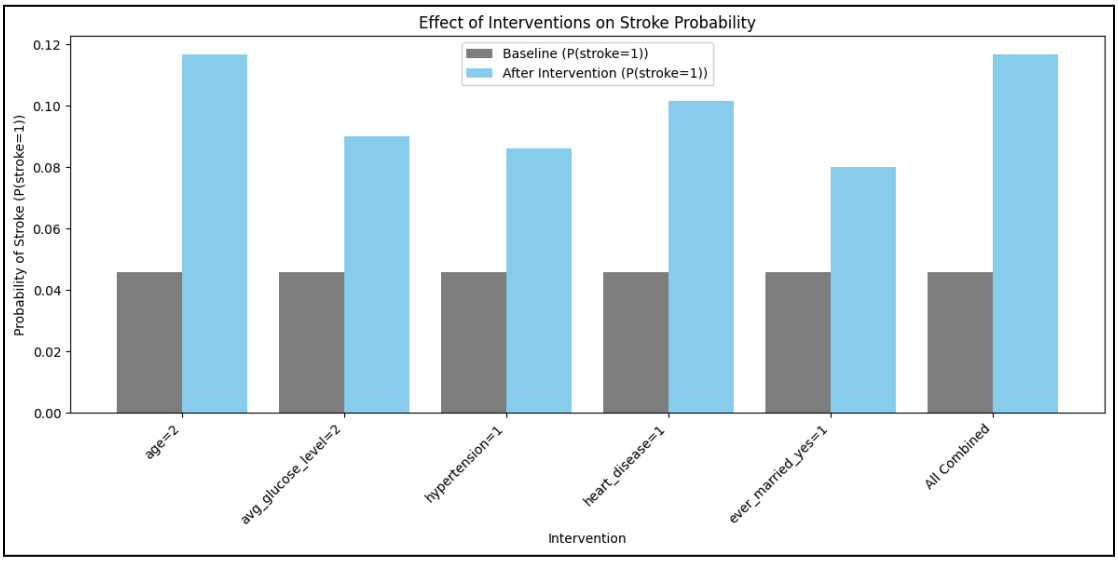
Gambar 10. Probabilitas Hasil Intervensi Variabel Parent

```
Intervention: P(stroke | age=2, avg_glucose_level=2, hypertension=1, heart_disease=1, ever_married_yes=1):
+-----------+---------------+
| stroke    |   phi(stroke) |
+===========+===============+
| stroke(0) |        0.8832 |
+-----------+---------------+
| stroke(1) |        0.1168 |
+-----------+---------------+
```

Gambar 11. Probabilitas Hasil Intervensi Variabel Terpilih



Gambar 12. Bar Chart Perbandingan Probabilitas Sesudah dan Sebelum Intervensi (1)



Gambar 13. Bar Chart Perbandingan Probabilitas Sesudah dan Sebelum Intervensi (2)

Gambar 11 dan Gambar 13 adalah gambar yang menunjukan bagaimana jika dilakukan intervensi kombinasi untuk seluruh variabel terpilih. Dari gambar tersebut ditemukan bahwa variabel avg_glucose_level, hypertention, ever_married_yes tidak memberikan perubahan yang signifikan. Ditemukan pula bahwa variabel yang paling memberikan perubahan signifikan adalah variabel age kemudian disusul oleh variabel heart_disease. Sehingga dapat disimpulkan bahwa variabel age dan heart_disease merupakan variabel parent dari variabel stroke.

5. Hasil

Dari analisis data maka didapatkan hasil :

a. Variabel Age merupakan parent utama dan mediator utama

Age adalah faktor risiko langsung utama stroke dengan korelasi tertinggi (0,25) dan efek kausal yang signifikan (11%). Variabel age menyebabkan peningkatan risiko stroke baik secara langsung maupun dengan mengkombinasi estimasi efek bersama dengan variabel lain seperti heart_disease dan hypertension. Namun dalam melakukan intervensi variabel age akan sulit dilakukan karena usia seseorang tidak bisa dilakukan intervensi pengurangan, sehingga langkah mitigasi lebih fokus pada mengelola efek lanjutan dari usia lanjut, seperti kontrol penyakit terkait usia.

b. Variabel Heart_Disease

Heart_Disease memberikan probabilitas tertinggi untuk stroke (17,03%), tetapi memiliki hubungan kausal tidak langsung terhadap stroke dengan melalui variabel age. Ini menunjukkan bahwa riwayat penyakit jantung pada pasien dengan usia tertentu (dalam hal ini usia lanjut) berkontribusi secara signifikan terhadap risiko stroke. Melakukan manajemen dan kontrol penyakit jantung, seperti penggunaan obat kardiovaskular bagi pasien dengan riwayat penyakit jantung, modifikasi gaya hidup, dan penanganan dini penyakit, menjadi langkah penting untuk mengurangi risiko stroke karena akan terjadi peningkatan probabilitas terjadi stroke sebesar 10,1% jika tidak dikelola dengan baik.

c. Variabel Hypertension merupakan penyumbang risiko tidak langsung

Meskipun variabel hypertension memiliki korelasi yang sama dengan heart_disease (0,13), dampaknya lebih sering dimediasi oleh variabel Age. Namun, probabilitas stroke akibat hipertensi masih tinggi (13,25%), menjadikannya variabel yang signifikan. Hipertensi adalah salah satu variabel yang lebih mudah dikelola melalui perubahan pola hidup (diet rendah garam, olahraga) dan terapi farmakologis, dan jika dilakukan intervensi hipertensi ini maka dapat menurunkan risiko stroke sebesar 8,6%.

6. Rekomendasi Eksperimen Lanjutan

Deteksi dini pada usia lanjut perlu dilakukan karena variabel usia memiliki peran sebagai mediasi untuk riwayat penyakit jantung dan riwayat hipertensi yang membuat peningkatan drastis untuk pasien dengan usia lanjut. Eksperimennya dapat berupa melakukan pengelolaan

secara agresif pada beberapa variabel risiko secara bersamaan (misalnya, pengelolaan hipertensi dan penyakit jantung pada usia lanjut) memiliki potensi dampak yang lebih besar dibandingkan hanya menangani satu variabel.

7. Rujukan

Aditya. Causal Modeling Hotel Booking Cancellation. 2021

Anas. Causal Modeling pada Klasifikasi Gaji. 202x

Daniel Dobrenz. Stroke - EDA, Model Tuning, and Predictions. 2025. https://www.kaggle.com/code/danieldobrenz/stroke-eda-model-tuning-and-predictions/notebook.

Fahmi. Causal Modeling Project Report. 2022.

Fedesoriano. Stroke Prediction Dataset. 2020. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data?select=healthcare-dataset-stroke-data.csv.

Maghfira. CDAG. 2025. https://github.com/035Maghfira/ProjectVisualAnalyticsforCausalModelingAnalysis