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

1. Ringkasan

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.

Variable	Description
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
```

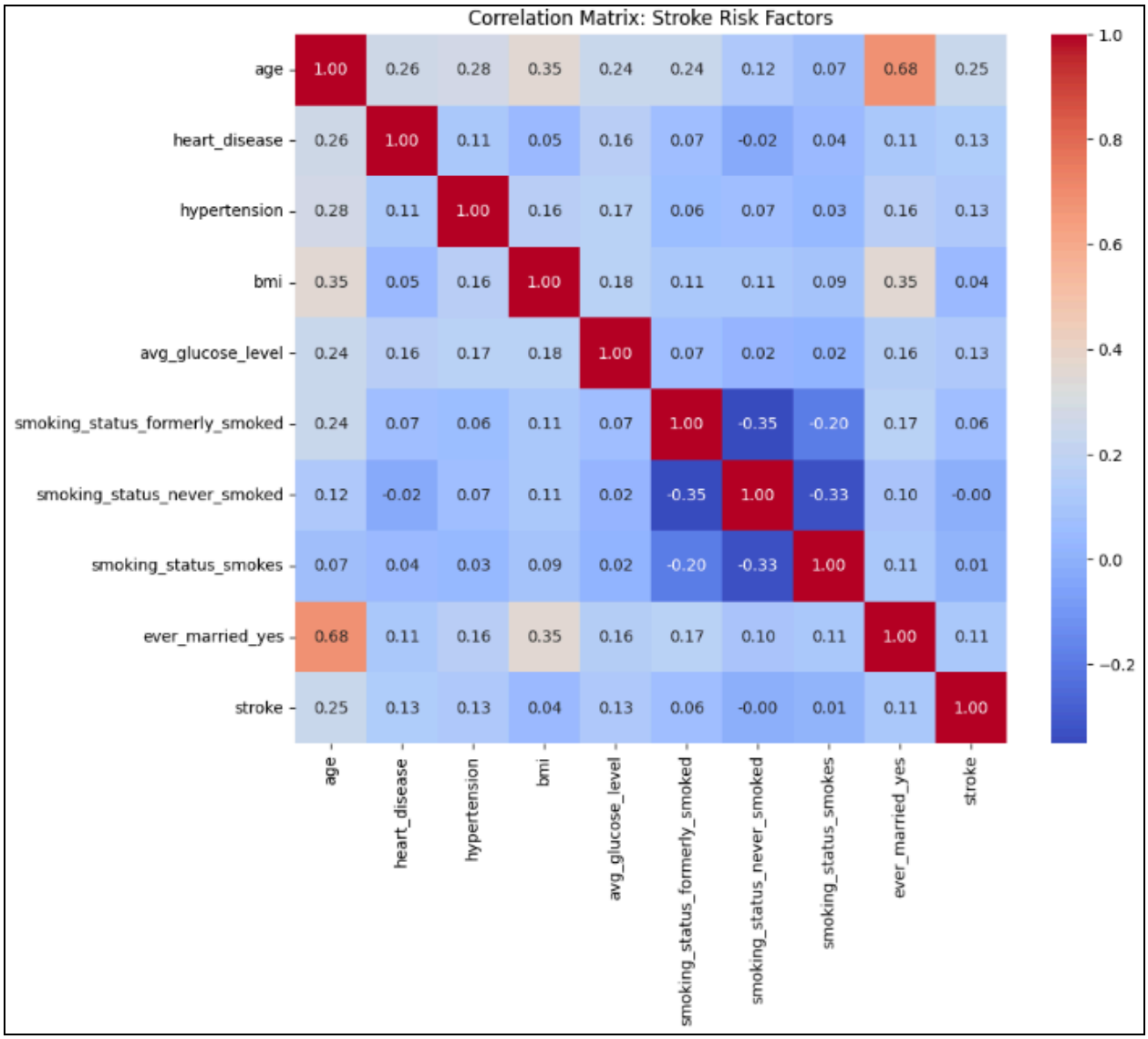
Gambar 1. Deskripsi Dataset Preprocessing

4. Analisis Data

   a. Eksplorasi data

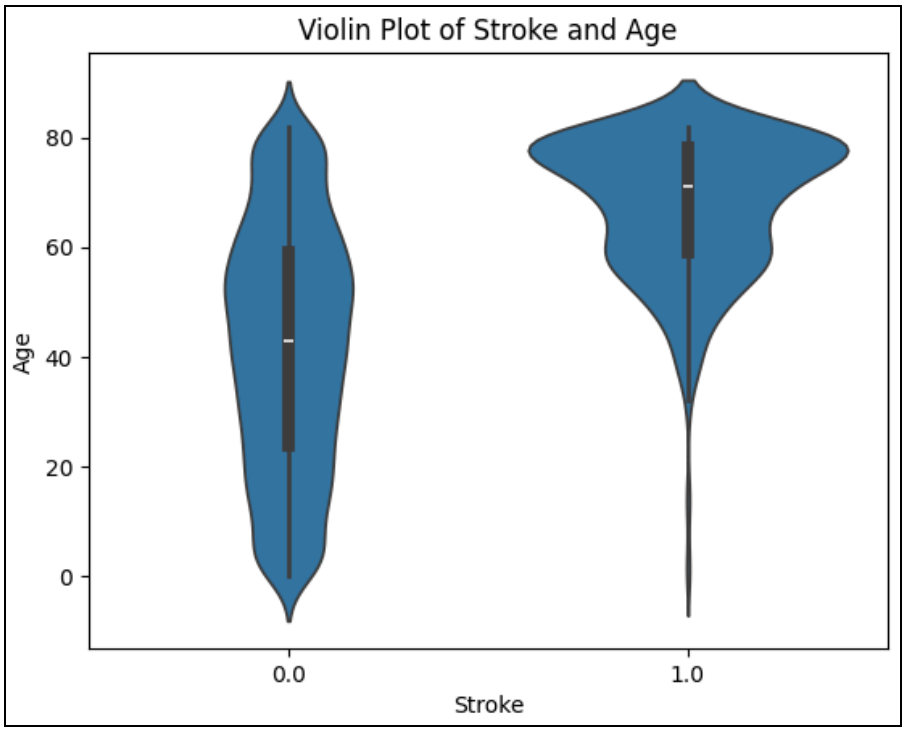
      Dataset stroke prediction ini telah mengalami preprocessing, sehingga dataset siap digunakan untuk proses analisis hubungan kausal. Pertama, kita perlu melihat korelasi antar variabel. Dari Gambar 2, yang merupakan heatmap korelasi antar variabel ini menunjukan bahwa variabel independen yang memiliki 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.

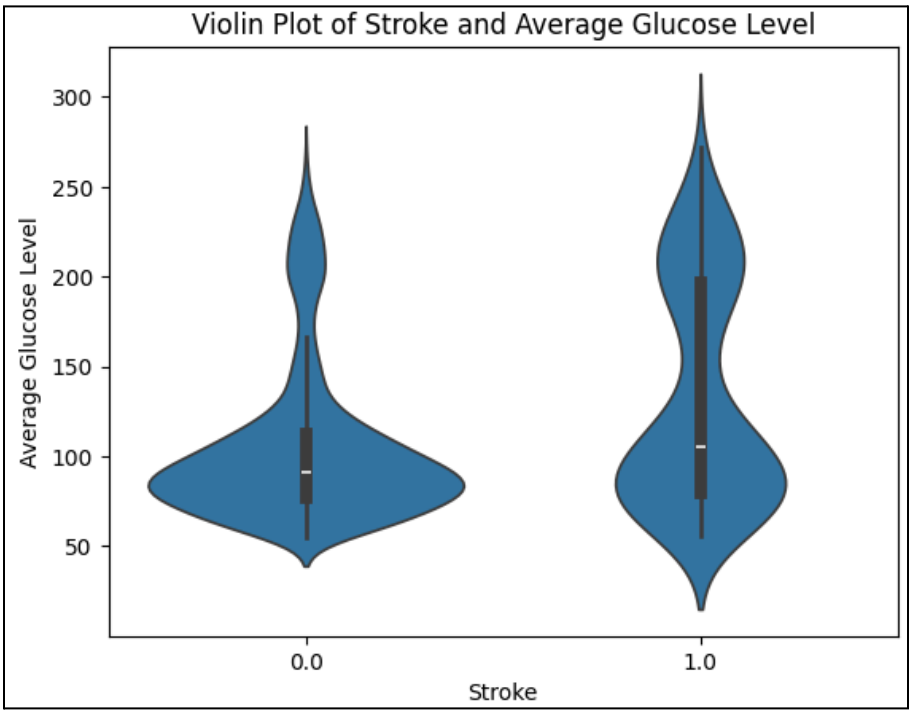


Gambar 2. Correlation Matrix antar variabel

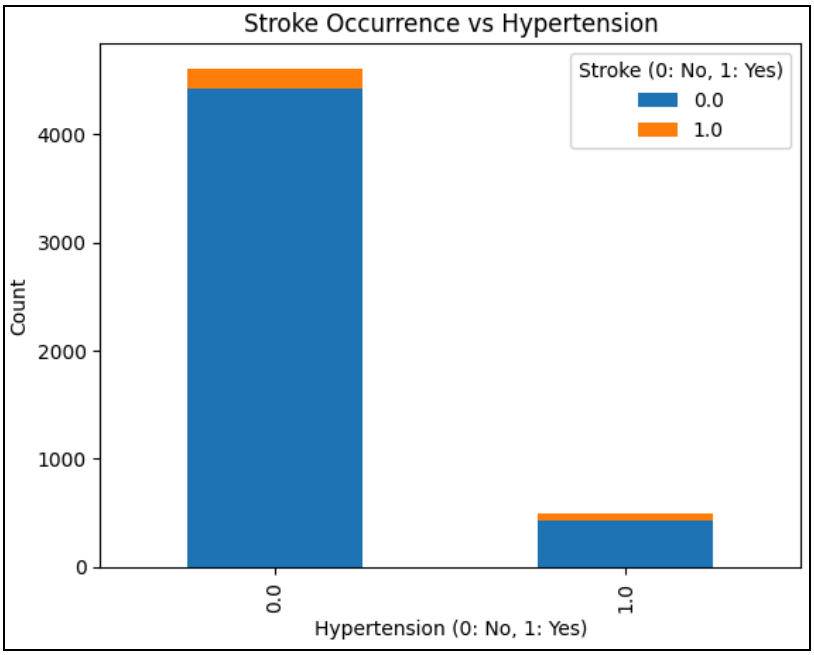
Gambar 3, 4, 5, dan 6 merupakan plot distribusi masing masing variabel dengan nilai korelasi tinggi (lebih dari 0.1) akan terjadinya stroke.



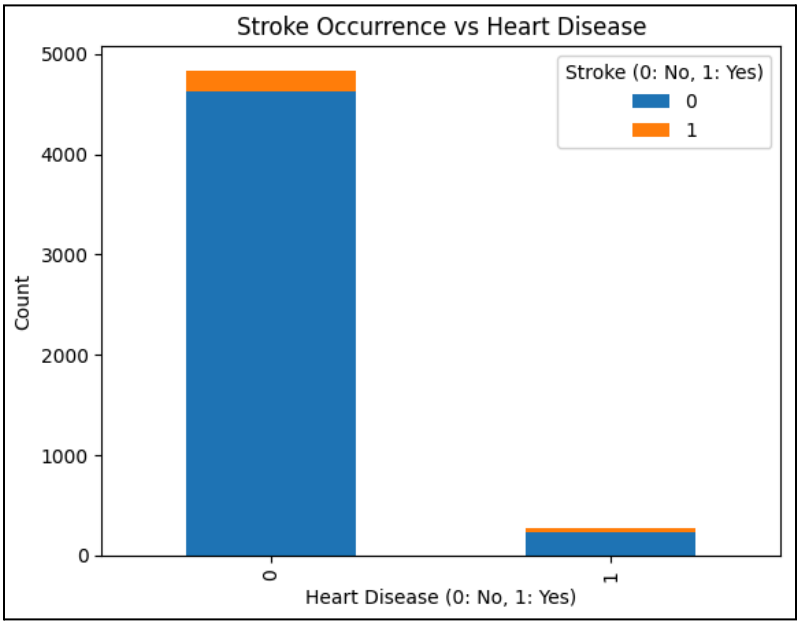
Gambar 3. Violin Plot Stroke Vs Age



Gambar 4. Violin Plot Stroke Vs Avg_Glucose_Level



Gambar 5. Stacked Bar Chart Stroke Vs Hypertension



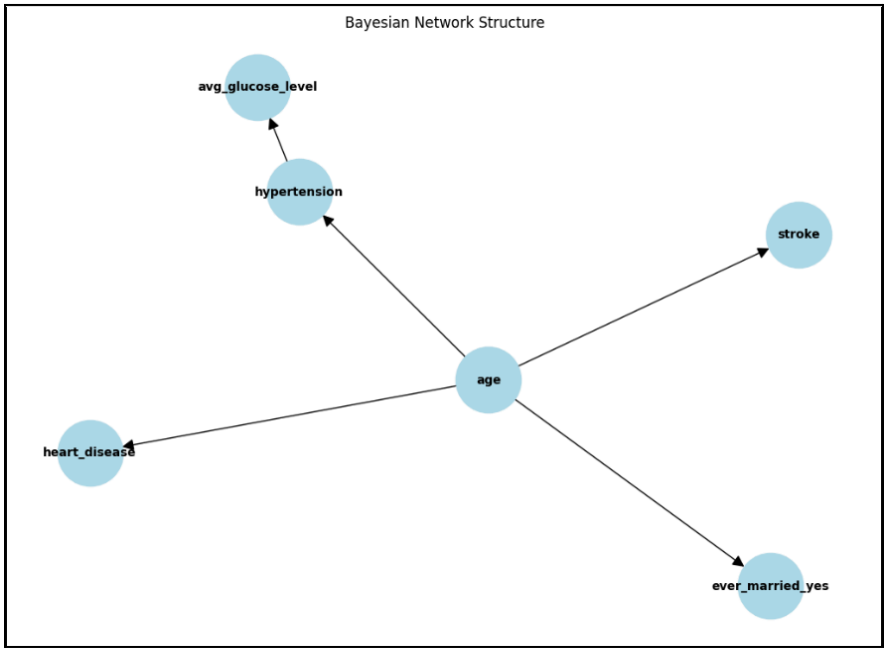
Gambar 6. Stacked Bar Chart Stroke Vs Heart_Disease

b. Pemilihan variabel penting

Dari korelasi tinggi ini, 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	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. Inference
d. Estimasi causal effect

5. Hasil

6. Rujukan

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

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