Nama : Gerrit Ezra Yudi Kairupan
NIM : 19/446777/TK/49882
Kelas : Rekayasa Data (Kelas A)

Laporan Rekayasa Data
Eksplorasi Data *Heart Disease*

A. Pendahuluan dan Deskripsi *Data Set*

Penyakit jantung adalah suatu kondisi saat jantung melangalami gangguan dalam berbagai bentuk, seperti gangguan pada pembuluh darah jantung, irama, katup, atau hal lain sebagainya. WHO mengumumkan bahwa penyakit dari kardiovaskular merupakan pembunuh nomor satu di dunia dimana terdapat tujuh belas juta orang meninggal pada tiap tahunnya terutama karena penyakit jantung. Maka dari itu kalimat "Mencegah lebih baik daripada mengobati" sangat relevan dengan keadaan sekarang karena jika dapat mengevaluasi risiko setiap pasien yang mungkin memiliki penyakit jantung, atau bahkan tidak hanya pasien tetapi semua orang dapat mencegah lebih awal penyakit tersebut.

Salah satu cara untuk mengevaluasi risiko yang dapat terjadi adalah dengan menganalisis data yang sudah ada sebelumnya. Terkait hal tersebut pada tugas kali ini, dibuat sebuah analisis eksplorasi data terkait penyakit jantung yang diperoleh dari data set UCI *Machine Learning*. Data yang akan digunakan adalah data set dari *Cleveland* karena data yang paling relevan dan disarankan oleh peniliti ML untuk dilakukan pemodelan analisis. Untuk data set dapat dilihat pada *clevelan.data-Heart Disease.csv*. Adapun deskripsi Data Atribut dari 303 *sample* data yang dianalisis sebagai berikut.

- age: Umur seseorang
- sex: Jenis Kelamin (1 = Laki-Laki, 0 = Perempuan)
- cp: Nyeri Dada yang dialami (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- trestbps: Tekanan Darah (mm Hg)
- chol: Kadar Kolesterol dalam mg/dl
- fbs: Gula Darah (> 120 mg/dl, 1 = true; 0 = false)
- restecg: Pengukuran elektrokardiografi (0 = normal, 1 = *having ST-T wave abnormality*, 2 = *showing probable or definite left ventricular hypertrophy by Estes' criteria*)
- thalach: Detak Jantung Maksimum
- exang: Angina yang diinduksikan (1 = yes; 0 = no)
- oldpeak: Depresi ST yang diinduksi oleh olahraga relatif terhadap istirahat
- slope: kemiringan segmen ST latihan puncak (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
- ca: The number of major vessels (0-3)
- thal: Kelainan darah yang disebut thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
- target: Penyakit Jantung (0 = no, 1 = yes)

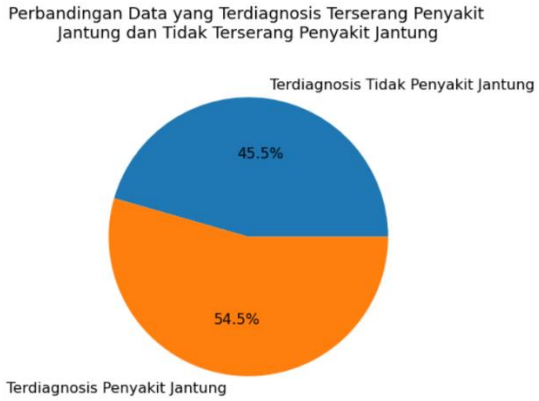
Diketahui bahwa dataset “target” adalah mereka yang mengalami penyakit jantung atau tidak. Untuk mengetahui hal tersebut sampel diuji dengan melihat besar penyempitan pada saluran darah jantung. Apabila <50% lebar penyempitannya maka dianggap tidak mengalami penyakit jantung (*value* = “0”) dan apabila >50% lebar penyempitannya maka dianggap mengalami penyakit jantung (*value* = “1”).

B. Isi dan Analisis Data

Eksplorasi analisis data menggunakan pemrograman bahasa Python pada Google Colab yang terlampir. Dalam eksplorasi data diperoleh hasil visualisasi data sebagai berikut.

1. Data yang terdiagnosis penyakit jantung

Pada grafik di bawah ini menunjukan banyaknya data yang terdiagnosis tidak mengalami penyakit jantung dan terdiagnosis penyakit jantung. Diperoleh bahwa dari 303 sampel yang ada terdapat 138 sampel (45.5%) tidak mengalami penyakit jantung dan 165 sampel (54.5%) terdiagnosis penyakit jantung. Dataset diperoleh bahwa pasien yang mengalami penyakit jantung lebih banyak sekitar 9% daripada pasien yang bebas dari penyakit jantung. Dari visualisasi data pada **Gambar 1** tersebut dapat disimpulkan bahwa sampel pasien dari *Cleveland* didominasi yang terserang penyakit jantung.



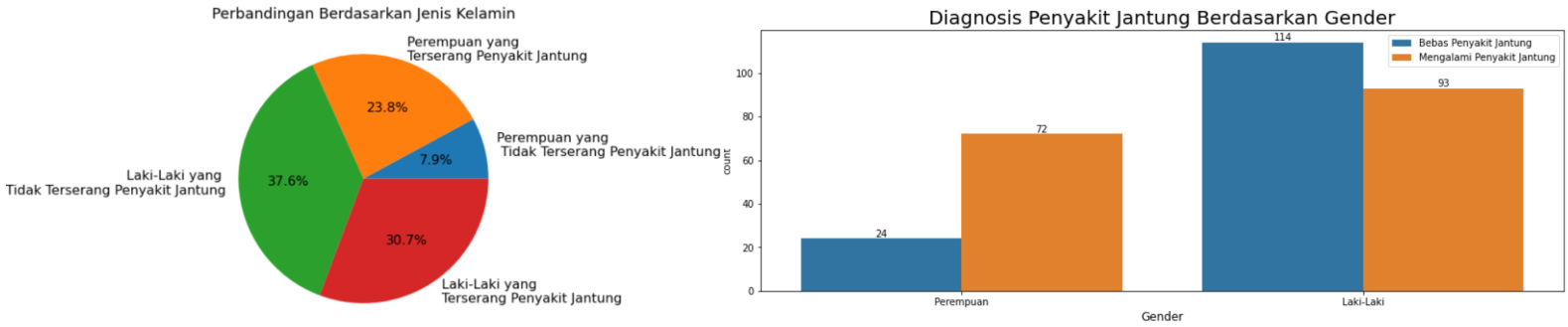

```
[11] df_target = df.groupby("target").size()
     df_target

     target
     0    138
     1    165
     dtype: int64
```

Gambar 1. *Informasi Data yang Terserang Penyakit Jantung*

2. Analisis data diagnosis penyakit jantung berdasarkan gender

Pada grafik di bawah ini menyajikan analisis data penyakit jantung berdasarkan Gender/ Jenis Kelamin. Perlu diketahui pada dataset bahwa gender perempuan memiliki nilai = “0” dan laki-laki memiliki nilai = “1”. Adapun analisis explorasi data grafik menggunakan bahasa pemrograman python ditujukan pada **Gambar 2**.

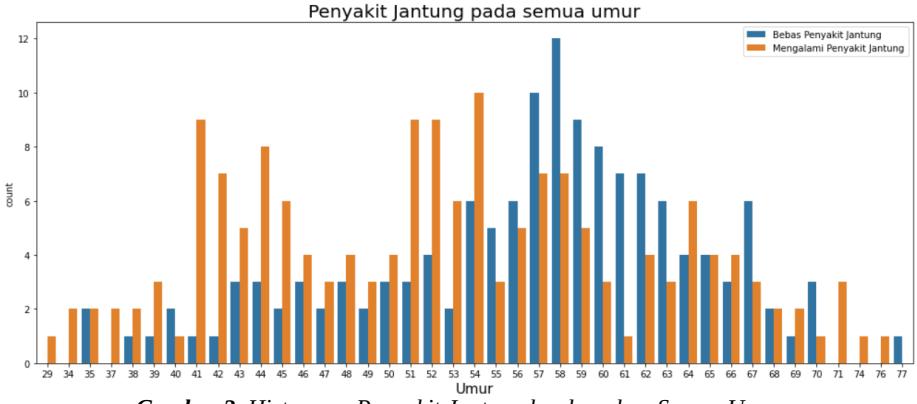


***Gambar 2**. Visualisasi Data Penyakit Jantung Berdasarkan Jenis Kelamin*

Pada grafik **Gambar 2** di atas menyajikan data persentase pengidap dan bukan pengidap penyakit jantung berdasarkan gender. Dapat diketahui dari visualisasi melalui *pie chart* dan Histogram di atas bahwa gender yang paling besar pengidap penyakit jantung adalah laki-laki, yakni sebesar 30,7% atau 93 Sampel pasien. Uniknya Gender yang paling sedikit mengidap penyakit jantung juga adalah gender laki-laki. Hal ini tentu jika hanya dilihat dari angka persentase dan jumlah saja dari data yang ada. Jika ditinjau dan dibandingkan antara pengidap laki-laki dan perempuan, gender perempuan lah yang memiliki jumlah selisih pengidap penyakit jantung yang paling banyak antara pengidap dan bukan. Pada kasus ini sampel perempuan masih tergolong sedikit dibandingkan laki-laki akan tetap jika ditinjau pada histogram di atas maka dapat dilihat bahwa pengidap penyakit jantung perempuan tergolong besar dibandingkan laki-laki. Dari data tersebut dapat ditarik kesimpulan bahwa gender perempuan memiliki potensi lebih besar terserang penyakita jantung dibandingkan gender laki-laki. Hal ini didukung oleh selisih antara pengidap dan tidak pengidap penyakit jantung lebih besar perempuan.

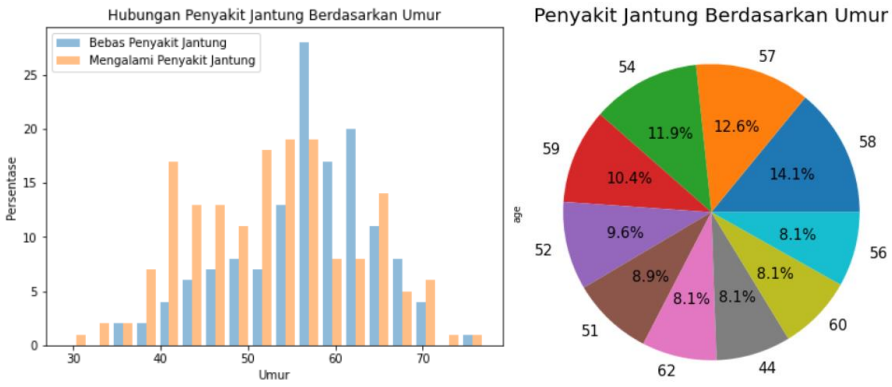
3. Analisis data diagnosis penyakit jantung berdasarkan umur

Pada grafik di bawah ini menyajikan analisis data penyakit jantung berdasarkan umur pasien. Pada analisis ini hanya menganalisis hubungan antara penyakit jantung dengan umur di luar variabel lainnya. Adapun visualisasi hubungan antara penyakit jantung dengan umur dari data sampel yang ada sebagai berikut.



***Gambar 3**. Histogram Penyakit Jantung berdasarkan Semua Umur*

Pada **Gambar 3** di atas merupakan histogram dari semua umur, dari histogram tersebut masih cukup sulit untuk dianalisis karena *range* data yang lebar/besar. Dari histogram tersebut diketahui pasien pengidap penyakit jantung berdasarkan sampel yang ada baling banyak pada umur 54 tahun, 52 tahun, 51 tahun, dan 41 tahun. Selain itu, tampak bahwa di atas umur 54 tahun pasien pengidap penyakit jantung memiliki jumlah yang relative cukup sedikit dibandingkan dengan 51-54 tahun. Meskipun demikian tampak juga bahwa umur 58 tahun memiliki jumlah paling besar yang bukan pengidap penyakit jantung. Hal tersebut memiliki indikasi bahwa pasien rumah sakit yang diambil sampelnya terdapat kemungkinan memiliki penyakit lain yang perlu dirawat di rumah sakit. Namun, pada umur 64 tahun pengidap penyakit jantung lebih tinggi dibandingkan yang bebas penyakit jantung. Untuk melihat range umur secara lebih jelas dapat ditujukan pada **Gambar 4** sebagai berikut.

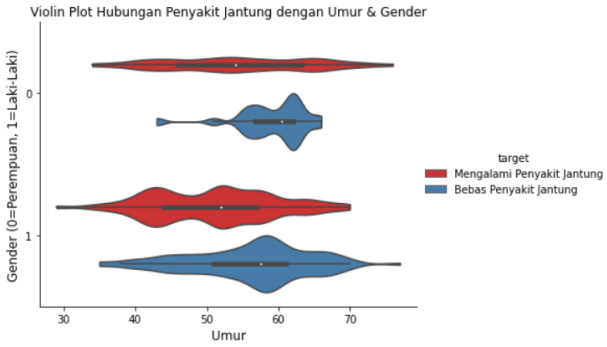


Gambar 4. *Range dan Persentase Penyakit Jantung berdasarkan umur*

Pada **Gambar 4** di atas dapat disimpulkan bahwa range umur yang paling beresiko terserang penyakit jantung adalah range umur 50-60 tahun dan range umur 40-50 tahun. Hal ini dudukung oleh persentase *pie chart* di atas dominasi umur pengidap penyakit jantung adalah pada umur 50-an dan terdpat umur 60 & 62 tahun dan 44 tahun. Analisis terhadap umur ini diluar variabel lainnya seperti gender, kalori, dan tekanan darah.

4. Analisis data diagnosis penyakit jantung berdasarkan gender dan umur

Pada **Gambar 2, Gambar 3,** dan **Gambar 4** dapat diketahui penyakit jantung dengan masing-masing variabel. Untuk mengetahui hubungan antara penyakit jantung dengan Umur dan Gender dapat dimodelkan menggunkan visualisasi data *Violin Plot* yang merupakan visualisais plot gabungan antara Box Plot dengan KDE (Estimasi Kepadatan Kernel). Adapun hasil visualisais menggunakan *Violin Plot* hubungan antara Penyakit Jantung dengan Umur dan Gender sebagai berikut.



Gambar 5. *Violin Plot Hubungan Penyakit Jantung dengan Umur & Gender*

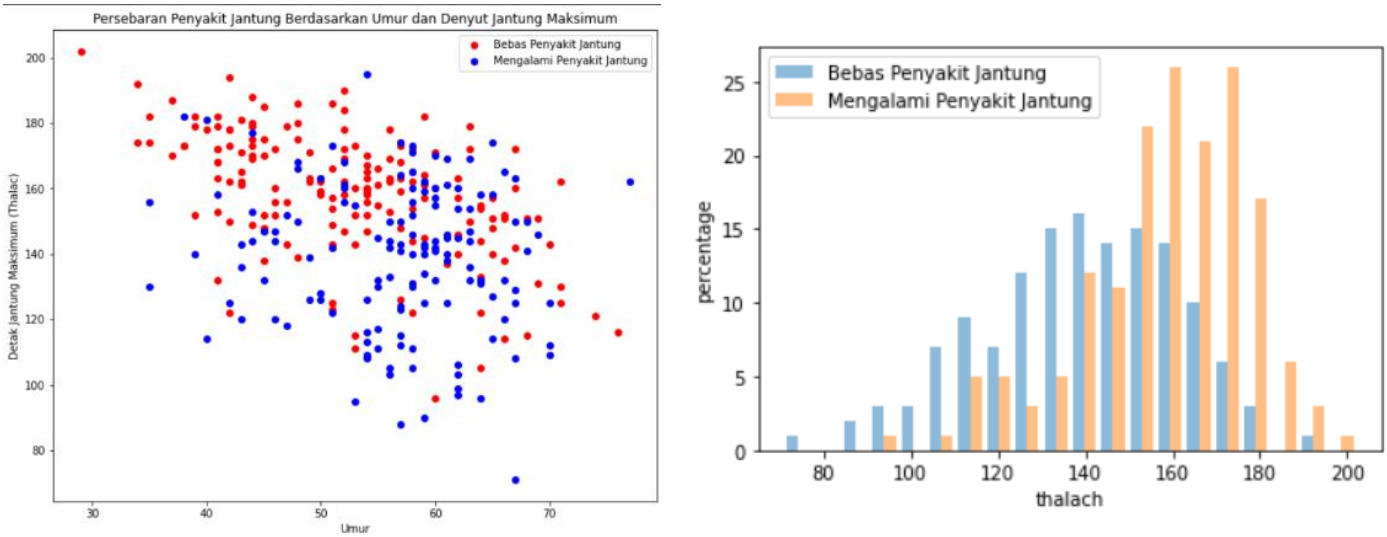
Pada **Gambar 5** di atas tampak merupakan *violin plot* pada sumbu x adalah umur dan sumbu y adalah gender (dimana *value* 0 adalaha perempuan dan *value* 1 adalah laki-laki). Dari *plot* tersebut diketahui bahwa penderita penyakit jantung perempuan (berwarna merah) merata dan melebar ke hampir semua umur serta memiliki puncak yang rata dibandingkan penderita penyakit jantung laki-laki yan memiliki banyak puncak tapi tidak merata. Pada penderita perempuan rata-rata pada sekitar 55 tahun dan sehat dari penyakit jantung rata-rata di sekitar umur 62 tahun. Untuk yang bebas dari penyakit jantung pada gender perempuan diketahui persebaran umur tidak merata dengan puncak-puncak tertinggi berada pada umur tertentu seperti mendekati umur 55 dan 62 tahun dibandingkan gender laki-laki yang sehat dari penyakit jantung memiliki persebaran umurnya merata dan memiliki puncak di sekitar 58 tahun. Dari *violin plot* ini pembaca dapat lebih paham persebaran-persebaran data yang terjadi khususnya gender laki-laki dan perempuan mengikuti umur pasien.

5. Analisis data diagnosis penyakit jantung berdasarkan umur dan detak jantung maksimum (thalach)

Pada **Gambar 5** dapat diketahui hubungan antara gender dengan umur untuk mengetahui sebaran penderita penyakit jantung. Analisis juga dapat dilakukan anatar umur dan detak jantung, adapun analisis visualisasi dilakukan menggunakan *scatter plot* pada **Gambar 6**.

Pada **Gambar 6** tampak bahwa persebaran data hubungan antara Detak Jantung Maksimum dengan Umur. Dapat diketahui bahwa mayoritas pasien yang mengalami penyakit jantung berada pada detak jantung maksimum di atas 140 ke atas dan titik sampel paling banyak berada pada*range* umur 50-60 tahun. Terdapat hal unik pada visualisasi data bahwa terdapat sampel yang berumur 30 tahun memiliki detak jantung maksimum di atas 200. Adapun keunikan yakni pada salah satu sampel memiliki *thalach* yang cukup tinggi di sekitar 190-an tetapi tidak mengalami penyakit jantung. Dengan

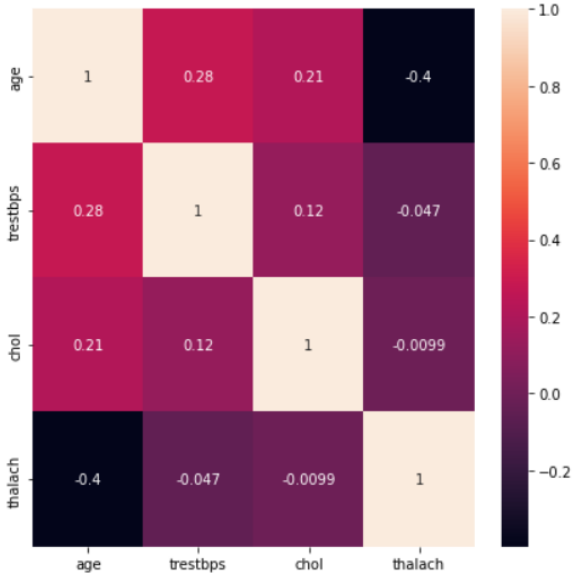
analisis data yang ada pada **Gambar 6** dapat ditarik kesimpulan bahwa hubungan antara denyut jantung dengan umur memiliki hubungan yang tidak menentu dimana tidak terdapat *range* umur tertentu memiliki denyut jantung yang tinggi.



Gambar 6. Scatter Plot Hubungan Penyakit Jantung dengan Umur dan Denyut Jantung

6. Korelasi data diagnosis penyakit jantung menggunakan *Heatmap*

Selain analisis-analisis di atas, data juga dapat dianalisis secara keseluruhan menggunakan *heatmap* untuk mencari data korelasi antar data-data yang tersedia. *Heatmap* menunjukkan intensitas relatif dari nilai yang ditperoleh untuk menetapkan setiap nilai representasi warna. Semakin mendekati 1 maka korelasi antar variabel lebih baik, sedangkan korelasi yang mendekati 0 bahkan negatif memiliki korelasi yang tergolong masih kurang. Adapun *heatmap* korelasi yang diperoleh ditunjukan pada **Gambar 7** sebagai berikut.



Gambar 7. *Heatmap Korelasi Data Penyakit Jantung*

Dari komponen-komponen data di atas dapat tampak korelasi-korelasi yang terjadi antar variabel yang ada. Dapat diketahui bahwa korelasi *thalach* atau Denyut Jantung Maksimum memiliki korelasi negatif pada semua komponen khsusnya pada komponen Umur. Seperti analisis sebelumnya diketahui bahwa korelasi *thalach* dengan umur cukup jauh karena denyut jantung maksimum tidak menentu pada berbagai umur. Sehingga dari data tersebut dapat ditarik kesimpulan juga bahwa detak jantung maksimum memiliki korelasi ataupun hubungan yang tidak menentu antara komponen-komponen lain. Berbanding dengan Komponen umur yang memiliki korelasi yang lebih tinggi dan lebih baik. Hal ini berarti umur dapat mempengaruhi peningkatan kolesterol dan tekanan darah.

C. Kesimpulan

Berdasarkan analisis yang dilakukan dapat diketahui bahwa korelasi terbaik dalam melakukan analisis dan prediksi adalah kolom/komponen Umur, hal tersebut seperti yang telah divisualisasikan sebelum-sebelumnya bahwa Umur memiliki hubungan yang cukup besar antar komponen lain dan dalam segi ilmu kesehatan dapat dihubungkan. Meskipun demikian, korelasi-korelasi yang dianalisis dan diperoleh tersebut masih tergolong kecil atau belum mendekati nilai 1, hal tersebut memiliki banyak faktor seperti distribus pemerolehan data yang tidak merata (seperti pada kasus jumlah pasien laki-laki lebih banyak dibandingkan pasien perempuan) ataupun jenis-jenis data yang diperoleh masih kurang varian untuk mencari korelasi lainnya. Maka dari itu, untuk memperoleh data yang lebih akurat dan presisi untuk penelitian alangkah baiknya data dapat kembali ditambah agar distribusinya merata dan komponen-komponen lainnya dapat ditambah untuk lebih mengeksplor data yang lebih dalam lagi.

REFERENSI

Hastono, Sutanto Priyo. (2016). *Analisis Data Pada Bidang Kesehatan*. Depok: PT. RAJAGRAFINDO PERSADA

Herho, Sandy Hardian Susanto. (2019). *Tutorial Visualisasi Data Menggunakan Seaborn.* Bandung: WCPL ITB.

Nimah, SupraptoYatim Lailun. (2019). *Pengantar Analisis Data Menggunakan Python*. Surabaya: Deepublish

Willy, dr. Tjin. (2018, Oktober 1). *Penyakit Jantung*. alodokter.com. https://www.alodokter.com/penyakit-jantung.