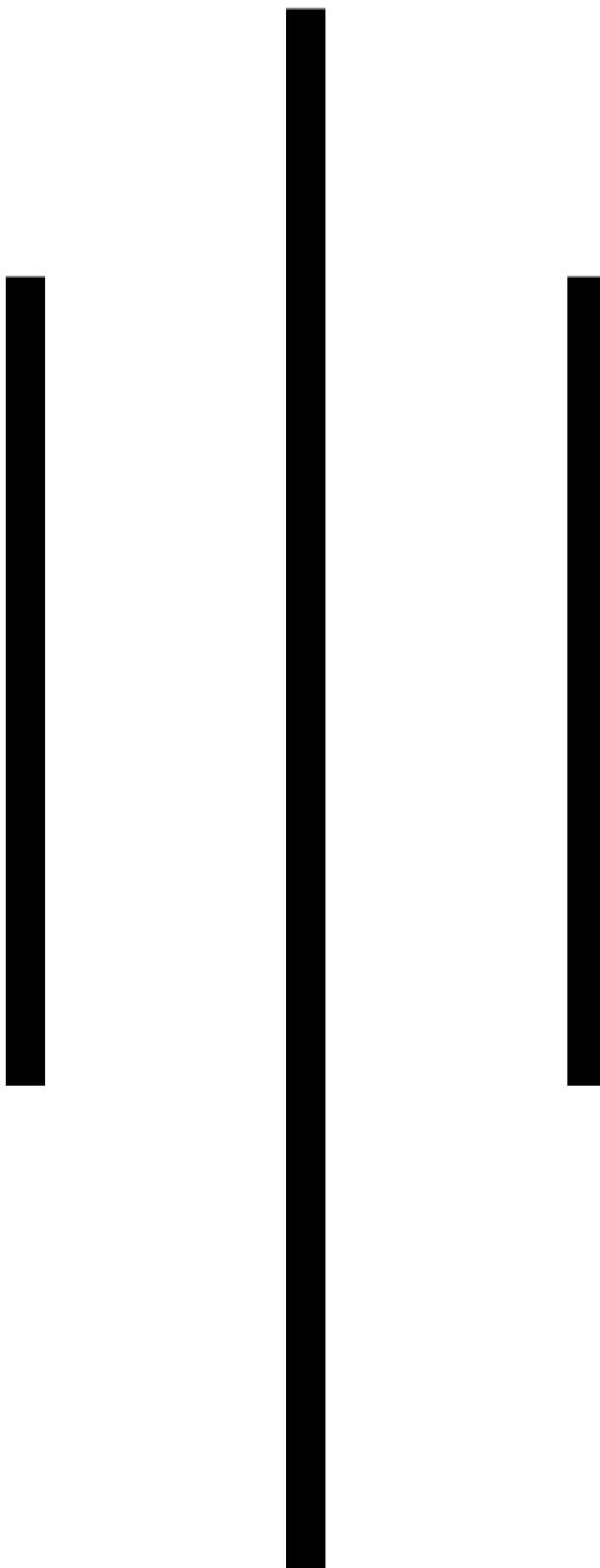


ANALISIS DAN PREDIKSI KADAR KOLESTEROL TOTAL SESEORANG DARI
VARIABEL-VARIABEL TERTENTU MENGGUNAKAN *MACHINE LEARNING*



Tim dataNoobs:

Salman Faiz Hidayat

Nabila Yumna Naafi'a

Hafid Sasayuda Ambardi

Daftar Isi

Daftar Isi.....	1
Bab I Pendahuluan.....	2
Latar Belakang.....	2
Rumusan Masalah.....	2
Tujuan.....	3
Bab II Pembahasan.....	3
Data Cleaning.....	3
Exploratory Data Analysis (EDA).....	3
Machine Learning Deployment.....	3
Hasil Prediksi.....	3
Bab III Penutup.....	3
Kesimpulan dan Saran.....	3
Daftar Pustaka.....	3

Bab I Pendahuluan

Latar Belakang

Kolesterol tinggi merupakan penyakit yang dialami oleh banyak orang di Indonesia. Persentase penderita kolesterol tinggi di Indonesia mencapai 28%, dan 7.9% meninggal dari penyakit ini (“Kolesterol”, n.d.). Parahnya lagi, kolesterol adalah babit dari penyakit-penyakit yang lebih mematikan, seperti penyakit jantung dan stroke. Dalam skala dunia, peningkatan kolesterol menyebabkan 2.6 juta kematian, dan menjadi penyebab kematian dini ataupun disabilitas di hari tua untuk 29.7 juta orang (“Raised cholesterol”, n.d.). Beberapa faktor yang sering dipertimbangkan ketika mengamati kadar kolesterol seseorang termasuk (“Causes and Risk Factors”, n.d.,):

- Gaya hidup tidak sehat (seperti konsumsi makanan kaya akan lemak jenuh)
- Usia
- Jenis Kelamin
- Keluarga atau keturunan
- Komorbiditas, atau pengidapan penyakit lain
- Pengobatan untuk penyakit lain
- Dan faktor-faktor lain seperti kegemukan, diabetes, ataupun komplikasi-komplikasi lain (“High Cholesterol”, n.d.,).

Maka dari itu, dengan melakukan analisis data menggunakan *machine learning*, faktor-faktor utama penentu kolesterol total seseorang dapat diketahui dan dipastikan secara lebih jelas. Dengan pengetahuan tersebut, seseorang bisa mengambil langkah-langkah pencegahan secara terarah sehingga kolesterol tidak bertambah dan menyebabkan penyakit-penyakit fatal.

Rumusan Masalah

Dengan mempertimbangkan latar belakang diatas, maka masalah yang dapat dirumuskan adalah:

1. Variabel-variabel (atau faktor) apa yang mempengaruhi kadar kolesterol total seseorang?
2. Apa korelasi antar variabel yang akan diolah?
3. Faktor-faktor apa yang paling berpengaruh terhadap kolesterol total?
4. Apa interpretasi yang dapat diambil dari pemodelan?

Tujuan

Faktor-faktor kesehatan apa yang paling berpengaruh terhadap nilai kolesterol total (CT).

Bab II Pembahasan

Data Cleaning

Data cleaning adalah proses ‘pembersihan’ atau perubahan format *data points* di dalam dataset (Wu, 2013). Ini dilakukan untuk membuat proses analisis menjadi lebih objektif (Chu et al, 2016), dan membuat model yang lebih akurat (Krishnan et al, 2016). Berikut adalah beberapa langkah yang diambil dalam proses *data cleaning*:

- Membuang data yang berduplikasi, untuk memastikan konsistensi pada model ketika membuat prediksi (Howe, 2001).
- Menghapus mayoritas row di mana kolom ‘Cholesterol Total (mg/dL)’ yang bernilai 187. Hal tersebut dilakukan untuk mencegah bias pada model, karena terdapat 785 baris dengan nilai 187 dari total 1339 baris. Dalam proses penghapusan baris tersebut, disisakan 20 baris dengan nilai 187 untuk mewakili baris-baris yang sudah terhapus.
- Menghapus *data points* yang memiliki nilai *outlier*, atau nilai yang berbeda jauh dari data yang lain (Aguinis et al, 2013). *Outlier* dibuang supaya model tidak terpengaruh secara ekstrim (Chatterjee et al, 1986). Berikut adalah karakteristik-karakteristik *outlier* yang didapat saat observasi:
 - ‘Tekanan darah (S)’ < 160.0
 - ‘Berat badan (kg)’ < 125.0
 - ‘IMT (kg/m²)’ < 40.0
 - ‘Glukosa Puasa (mg/dL)’ < 200.0
 - ‘Trigliserida (mg/dL)’ < 500
 - Nilai ‘Lingkar perut (cm)’ yang sama dengan nilai minimum kolom tersebut.
- Membuang beberapa kolom yang tidak digunakan:
 - ‘Responden’
 - Kolom ‘Responden’ hanya berfungsi sebagai penomoran data dan tidak berpengaruh pada prediksi.
 - ‘Tempat lahir’
 - Untuk menurunkan granularitas tempat lahir yang memiliki 170 nilai unik, ‘Tempat lahir’ dikelompokkan dan diubah menjadi kategori biner. Jika tempat lahirnya adalah wilayah metropolitan berdasarkan Universitas Sains dan Teknologi Komputer (2024), maka akan bernilai 1, dan 0 jika tidak. Uji ANOVA dapat digunakan pada data kategori untuk mengukur apakah terdapat perbedaan yang signifikan secara statistik pada nilai ‘Total Kolesterol (mg/dL)’ untuk setiap kategori tempat lahir. Dengan menggunakan uji ANOVA, kita dapat menghitung bahwa *p-value* dari tempat lahir biner adalah 0,513867. Artinya, terdapat 51% kemungkinan *null hypothesis* benar. Dalam kasus ini, *null*

hypothesis berarti tidak ada hubungan yang signifikan antara tempat lahir dengan kolesterol total. Oleh karena itu, kami tidak akan menyertakan fitur ini dalam model pembelajaran mesin. (Uji ANOVA juga dilakukan untuk data kategori ‘Jenis Kelamin’ dan ditemukan *p-value* 0.048966 jadi kemungkinan *null hypothesis* benar kurang dari 5%).

- ‘Tinggi badan (cm)’ dan ‘Berat badan (kg)’
 - Kedua variabel ini telah diwakilkan oleh ‘IMT (kg/m²)’ atau *body mass index*.
- Membuang nilai modus pada kolom ‘Fat’ dan ‘Visceral Fat’ karena menyebabkan kejanggalan dalam hubungan dengan variabel lain. Hal tersebut akan menghasilkan nilai *Nan* yang akan diisi dengan *KNN Imputer*.

Selain pembersihan data, dataset ditransformasi, sehingga model bisa memproses variabel-variabel data dengan lebih baik dan mendapat akurasi yang lebih tinggi (Khurana et al, 2017). Proses ini disebut dengan *feature engineering*. Dalam kasus ini, *feature engineering* dilakukan dengan langkah-langkah berikut:

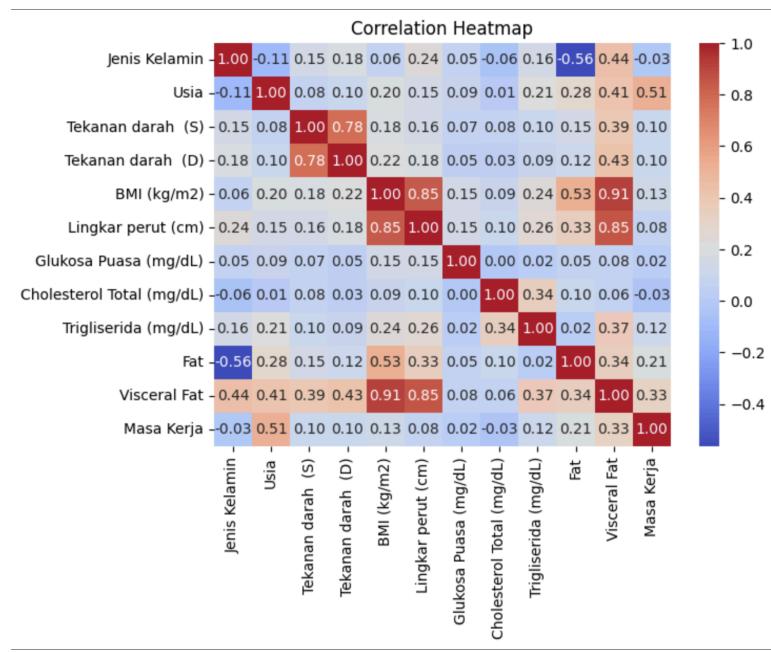
- Melakukan *encoding* pada kolom ‘Jenis Kelamin’ (Male (M): 1, Female (F): 0).
- Mengambil nilai logaritma natural dari kolom ‘Glukosa Puasa (mg/dL)’ dan ‘Trigliserida (mg/dL)’
- Memproses data *training* dan *test* secara terpisah untuk menghindari *data leakage*:
 - Data *training* digunakan untuk melatih model.
 - Data *test* digunakan untuk menguji kualitas model setelah dilatih.
 - *Scaling* data menggunakan *Standard Scaler*.
 - *Scaling*, atau merubah jangkauan data (Cook, n.d.,), dilakukan agar algoritma KNN imputation bisa lebih mudah mengidentifikasi “tetangga” dari data yang kosong sehingga imputasi bisa dilakukan dengan lebih baik.
 - Melakukan imputasi data yang hilang dengan *KNN Imputation* pada kolom ‘Fat’ dan ‘Visceral Fat’ (*neighbors* = 3).
 - Mengembalikan label kolom dataset yang hilang pada tahap *scaling*.

Exploratory Data Analysis (EDA)

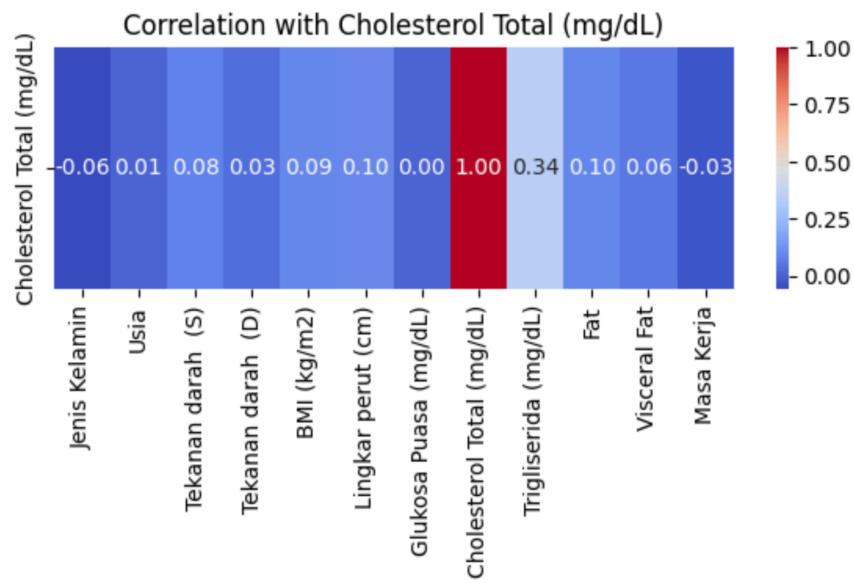
Exploratory Data Analysis (EDA) adalah tahap dimana karakteristik data dieksplorasi (Morrison, 2010), dan pola-pola dalam data menggunakan komputer (Behrens, 1997) . Intinya, proses ini berfokus untuk mengetahui informasi yang bisa diungkap oleh data (Morgenthaler, 2018). Berikut adalah beberapa visualisasi data yang telah dibuat untuk EDA:

Heatmap

Heatmap adalah alat visualisasi data yang digunakan untuk menganalisis pola-pola yang ada diantara variabel-variabel (Gu et al, 2021). Berikut adalah visualisasi *heatmap* dengan lengkap dengan matriks korelasi antar variabel untuk data 2:



Untuk mengetahui korelasi variabel lain dengan kolesterol total, bagian yang perlu diobservasi adalah barisan yang dilabeli ‘Cholesterol Total (mg/dL)’ pada sumbu Y grafik:

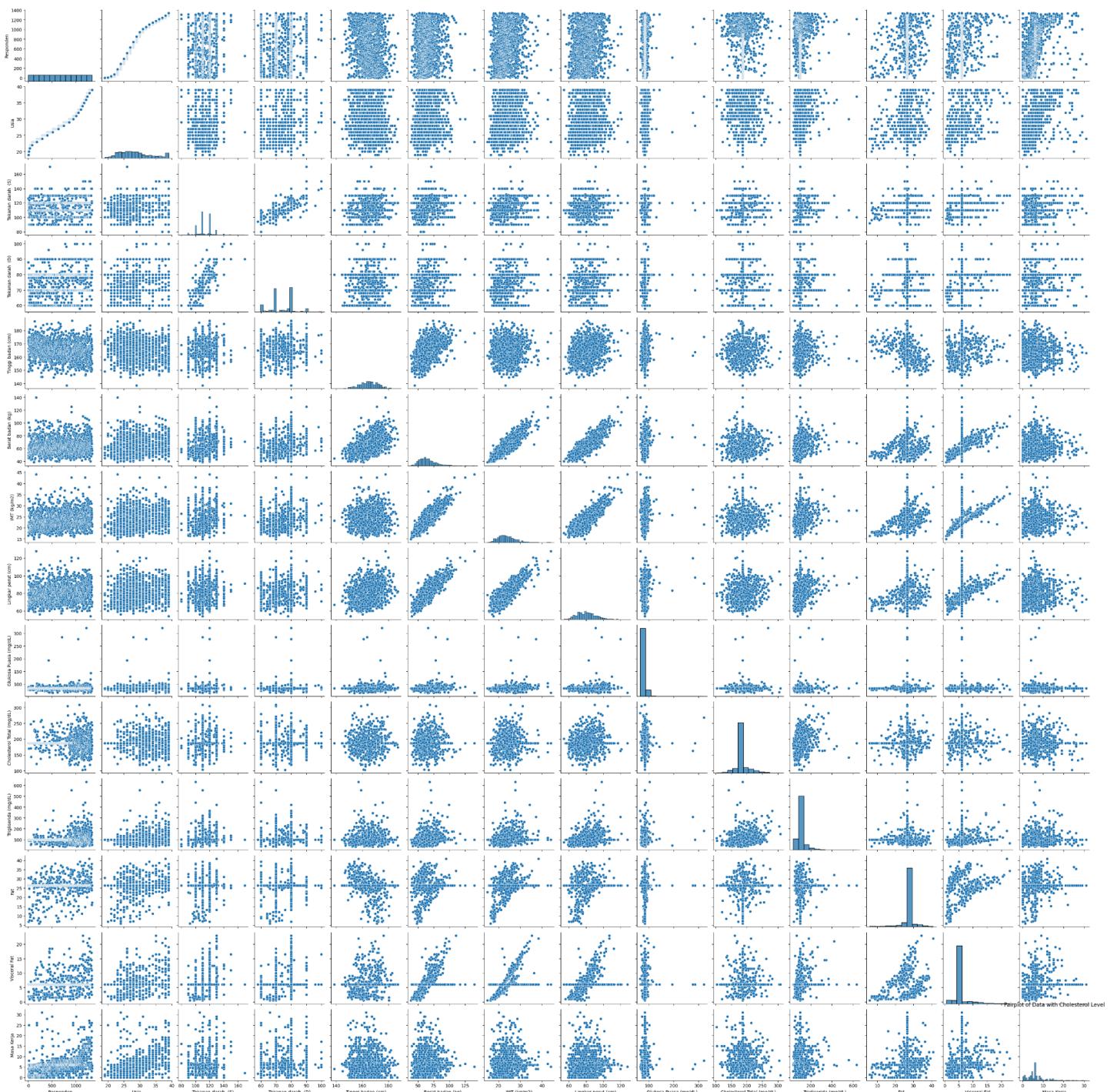


Korelasi yang memiliki nilai positif menunjukkan bahwa semakin tinggi nilai variabel pada sumbu X, semakin tinggi ‘Cholesterol Total (mg/dL),’ sedangkan korelasi dengan nilai negatif menunjukkan bahwa semakin tinggi nilai sumbu X, semakin rendah ‘Cholesterol Total (mg/dL)’ (Chip, 2024).

Menurut grafik ini, hampir semua variabel memiliki korelasi yang kecil, yaitu dibawah atau sama dengan 10%, baik positif maupun negatif.

Pengecualian untuk kolom ‘Trigliserida (mg/dL)’ yang memiliki korelasi positif sebesar 34%. Korelasi ini setuju dengan sebuah studi yang mengatakan bahwa ada asosiasi yang kuat antara trigliserida dan kolesterol total yang berada dibawah 250 mg/dL (Freedman et al, 1988). Akan ditunjukkan bahwa trigliserida akan menjadi variabel yang paling berpengaruh dalam prediksi model *machine learning* pada seksi Machine Learning Deployment, baik secara regresi (prediksi numerik) maupun klasifikasi (prediksi kategorik).

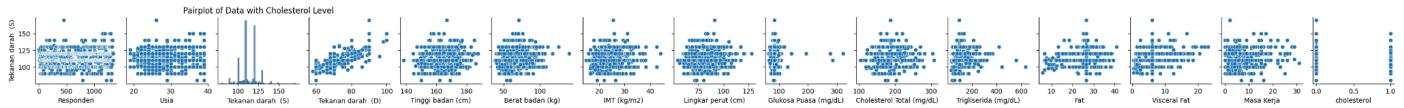
Pairplot



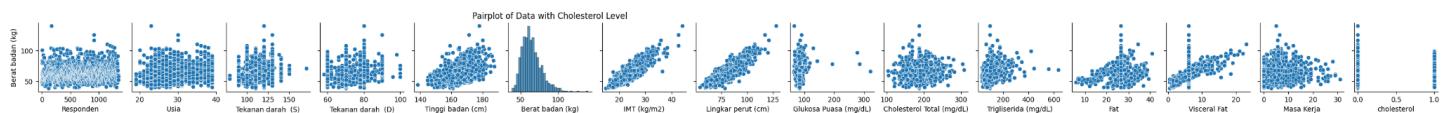
Pairplot adalah grafik yang menunjukkan kombinasi antara variabel-variabel numerik dan kategorik. Ilustrasi ini dapat memberi pengetahuan akan struktur data yang memiliki multivariabel (Emerson et al, 2013).

Selain itu, *pairplot* dapat memberikan kemudahan dalam menentukan *outliers*, dan mencari kejanggalan yang berpotensi mengganggu prediksi.

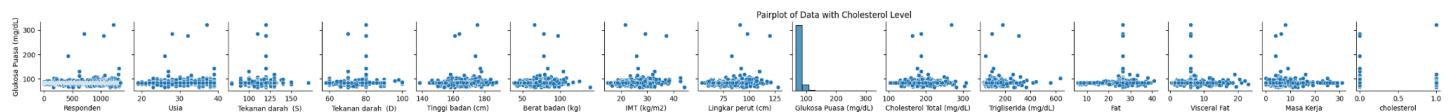
Pairplot digunakan dalam menganalisis outlier secara garis besar, berikut adalah potongan outlier dan data frame untuk modelling:



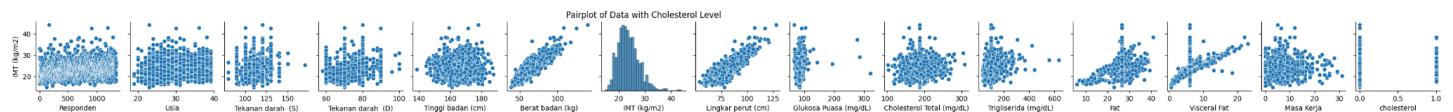
- Terdapat baris yang memiliki Tekanan Darah Sistol diatas 160.



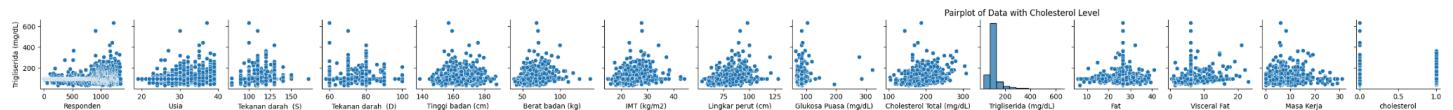
- Terdapat baris yang memiliki Berat badan diatas 125 kg.



- Terdapat baris yang memiliki Glukosa Puasa diatas 200 mg/dL.

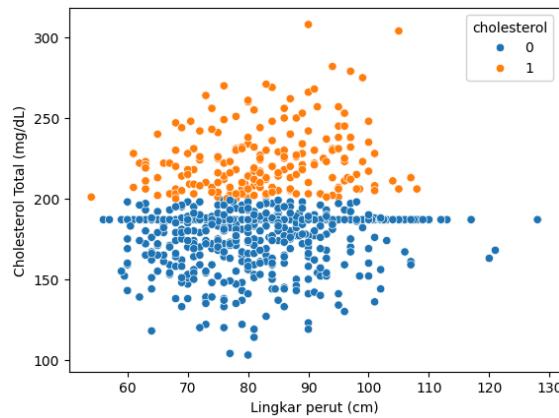


- Terdapat baris yang memiliki IMT/BMI diatas 40.



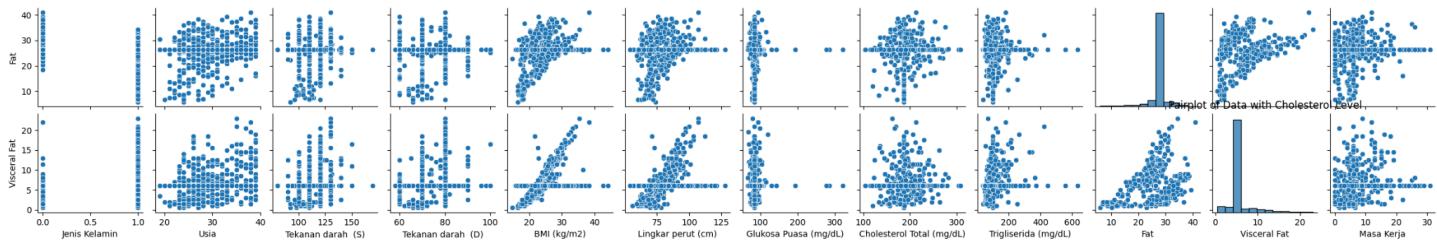
- Terdapat baris yang memiliki Triglycerida diatas 500.

Untuk pendekatan secara klasifikasi (yang akan dibahas lebih lanjut dibawah) terdapat tambahan outlier sebagai berikut:



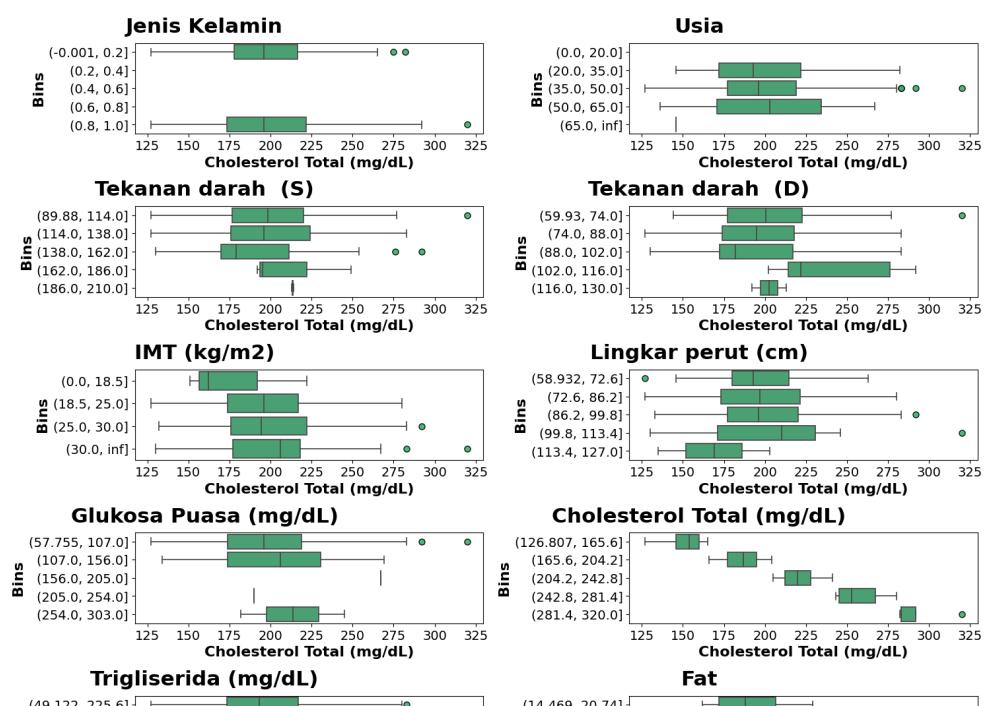
- Terdapat baris yang memiliki lingkar perut terkecil, tetapi terindikasi mempunyai kolesterol.

Selain *outlier*, terdapat kejanggalan pada data, utamanya pada kolom ‘Fat’ dan ‘Visceral Fat.’ Menurut observasi, terdapat satu nilai tertentu pada kedua kolom tersebut yang “mengganggu” korelasi yang tergambar. Gangguan tersebut berbentuk titik nilai yang tergambar tegak lurus dari sumbu X ataupun Y (tergantung referensi. Pada ilustrasi, tegak lurus terhadap sumbu Y):

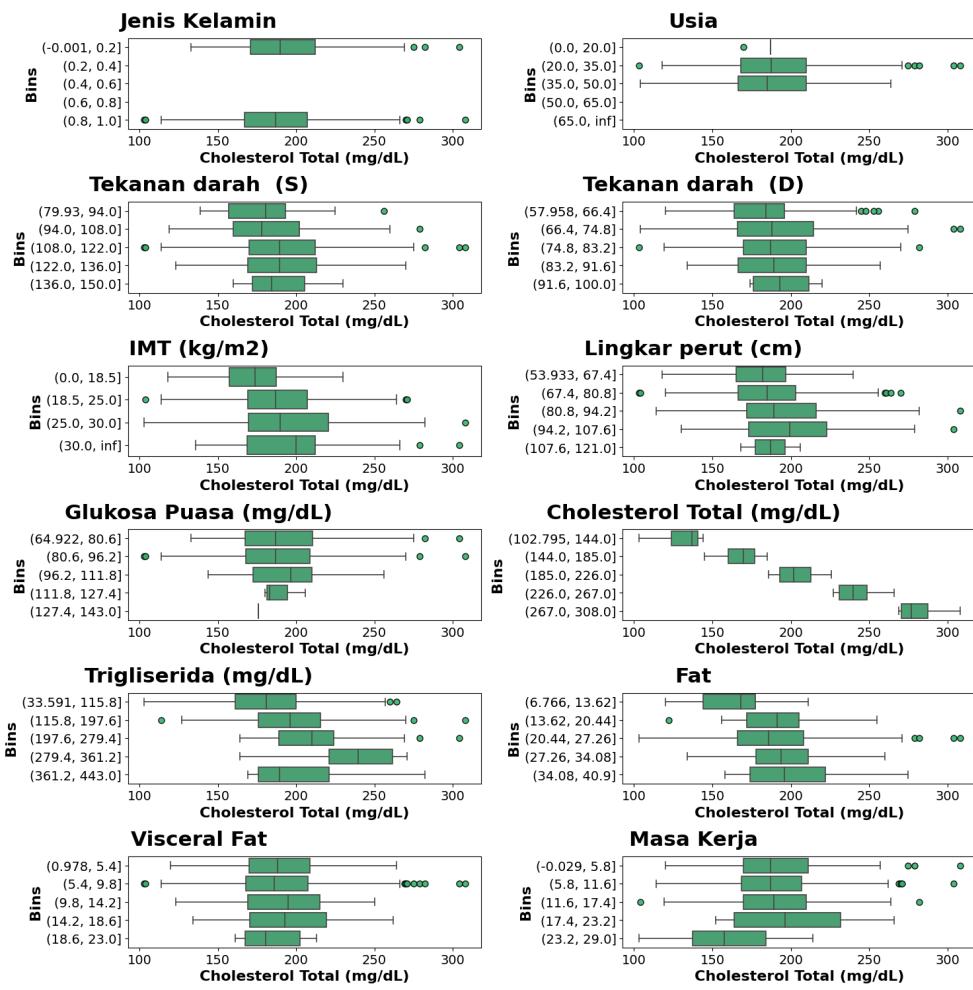


Setelah diobservasi, nilai tersebut merupakan nilai modus dari kolom ‘Fat’ dan ‘Visceral Fat.’ Dapat diasumsikan bahwa ini akan merusak korelasi natural data. Maka dari itu, modus ‘Fat’ dan ‘Visceral Fat’ dibuang dan dilakukan imputasi *KNN* (*K-Nearest Neighbors*), supaya korelasi data lebih alami. Kemudian, ini juga bermanfaat untuk modelling, karena model tidak akan menjadi *bias*, atau membuat prasangka yang salah, yang disebabkan oleh terlalu banyaknya data dengan nilai modus tersebut.

Boxplot Data 1



Boxplot Data 2



Grafik *Boxplot* adalah grafik yang menunjukkan distribusi data menggunakan kuartil.

Beberapa informasi penting dapat ditarik dari grafik *boxplot*:

- Kolesterol total yang rendah cenderung berasosiasi dengan masa kerja yang panjang.
- Bin ‘Visceral Fat’ dan ‘Lingkar perut (cm)’ tertinggi cenderung memiliki kolesterol total yang lebih rendah.
- ‘IMT (kg/m²)’ tinggi cenderung terasosiasi dengan kolesterol total yang lebih tinggi.

Machine Learning Deployment

Untuk modelling, dilakukan 2 pendekatan:

- Pendekatan pertama adalah prediksi secara regresi, atau memprediksi angka pasti kolesterol total.
- Pendekatan kedua adalah prediksi secara klasifikasi, atau memprediksi apakah kolesterol total seseorang tinggi atau rendah (Tinggi: ≥ 200 , rendah: < 200). Lalu, kolesterol tinggi di-*encode* dengan angka 1, sedangkan kolesterol rendah dengan angka 0. Sehingga, prediksi bersifat biner (*True/False*).

Extreme Gradient Boost Regressor (XGBRegressor)

XGBoost adalah algoritma *machine learning* yang bekerja berdasarkan metode *Gradient Boosting*, yang bekerja seperti berikut (Cook et al, n.d.,):

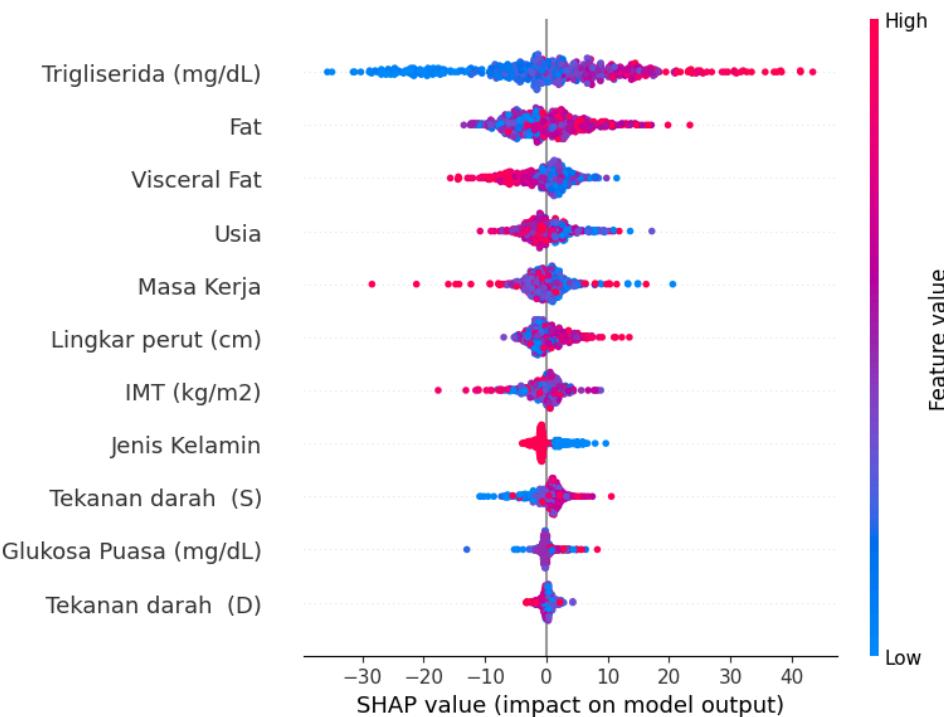
1. Model awal membuat prediksi secara naif.
2. Prediksi tersebut akan diukur menggunakan fungsi kerugian (*loss function*) yang mengukur error prediksi.
3. Berdasarkan fungsi kerugian, model baru akan dilatih, dan dikombinasikan dengan model awal untuk membuat model ansambel, dengan harapan menurunkan fungsi kerugian.
4. Kemudian dilakukan selama berulang kali, sesuai dengan parameter *n_estimator* (Brownlee, 2020).

Model ini digunakan karena model bisa bekerja bahkan pada data yang tidak seimbang, sambil diiringi dengan metode *resampling* yang baik (Zhang et al, 2022), seperti *upsampling* (menambah sample) dan *undersampling* (mengurangi sample).

Parameters yang digunakan untuk model XGBoost:

- *n_estimators* adalah 500
- *learning_rate* adalah 0,01
- *scale_pos_weight* adalah 3

Berikut adalah interpretasi hasil pembelajaran model XGBRegressor menggunakan SHAP (SHapley Additive exPlanation) (“Shapley Values”, 2023):



Diatas adalah *SHAP Summary Plot*, yang berfungsi untuk menunjukkan seberapa banyak sebuah variabel dipertimbangkan oleh sebuah model. Semua variabel diurut dari atas ke bawah, berdasarkan tingkat pengaruh variabel terhadap prediksi. Titik yang memiliki warna kemerahan menunjukkan bahwa nilai dari data tersebut memiliki nilai yang tinggi, dan titik yang berwarna kebiruan memiliki nilai yang rendah.

Sumbu X di bagian bawah grafik menunjukkan impact seperti apa yang dibuat oleh setiap nilai, utamanya secara positif atau negatif:

- Sebelah kanan sumbu X menunjukkan bahwa kolesterol total yang diprediksi akan bertambah, dengan data yang berkoresponden dengan bagian sumbu tersebut.
- sebelah kiri sumbu X menunjukkan bahwa kolesterol yang diprediksi akan berkurang, dengan data yang berkoresponden dengan bagian sumbu tersebut.

(Tripathil, 2023)

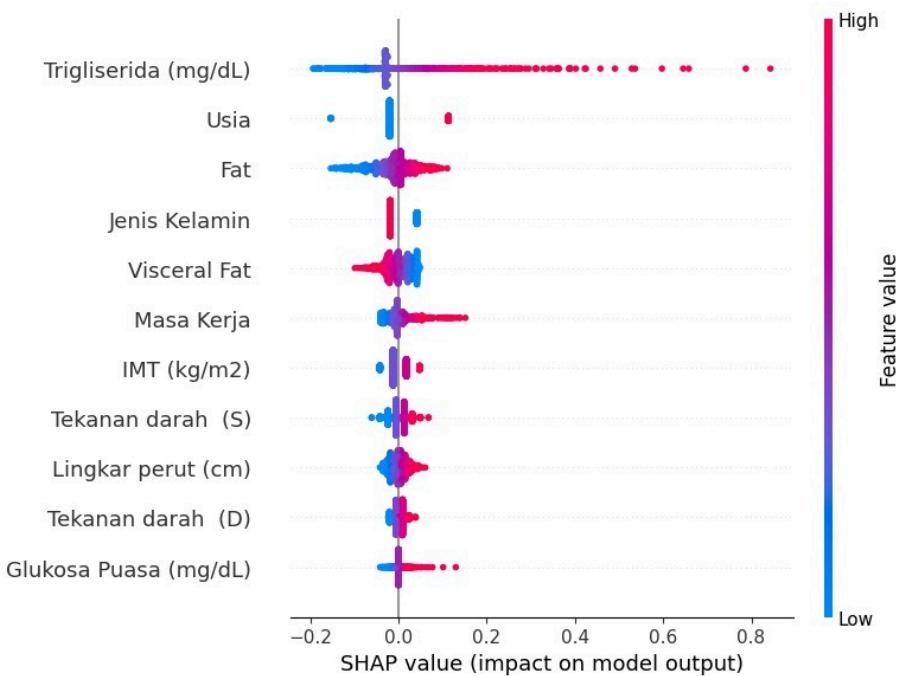
Perlu dicatat bahwa ‘Tempat lahir’ tidak ada, karena menurut kalkulasi *P-Value*, kolom tersebut tidak masuk ke dalam batasan nilai, yang membuat kolom tersebut signifikan secara statistika. Berikut adalah interpretasi objektif yang penting dari ilustrasi diatas:

- 3 prediktor terkuat untuk memprediksi kolesterol total adalah trigliserida, lemak (‘Fat’), dan lemak jahat (‘Visceral Fat’).
- Trigliserida adalah prediktor yang paling kuat. Kolesterol total dan trigliserida memiliki korelasi positif.
- Perempuan berusia muda rentan memiliki kolesterol tinggi.
- Semakin lama masa kerja seseorang, semakin rendah total kolesterol.
- Indeks Massa Tubuh (IMT kg/m²) tinggi tidak selalu memiliki kolesterol total yang tinggi. Disini ditunjukkan bahwa orang-orang dengan IMT tinggi memiliki probabilitas yang kurang lebih sama, untuk memiliki kolesterol total rendah maupun tinggi.
- Tekanan darah sistol (‘Tekanan darah (S)’) berkorelasi positif dengan kolesterol total, sedangkan tekanan darah diastol (‘Tekanan darah (D)’) berkorelasi negatif dengan kolesterol total.

Extreme Gradient Boost Classifier (XGBClassifier)

Algoritma model ini mirip atau sama dengan XGBRegressor, akan tetapi nilai yang diprediksi adalah diskret, bukan kontinu (Lev, 2022). Model ini sangat berguna untuk dataset yang diberikan, karena dapat mengatasi ketidakseimbangan di dalam data. Parameter ‘scale_pos_weight’ sangat berguna untuk menyeimbangkan prediksi kategori data, karena ini akan mendorong model untuk melakukan koreksi berlebih terhadap kelas yang menjadi minoritas (Kathula, 2023). Sehingga, model masih membuat prediksi terhadap kategori minoritas, meskipun data tidak mendukung.

Perlu dicatat bahwa terdapat perbedaan dalam pengolahan data. Kolom ‘Usia’ dan ‘IMT (kg/m²)’ dilakukan *ordinal encoding*, atau encoding menggunakan bilangan bulat. Ini dilakukan karena model mendapat akurasi yang lebih baik dengan *encoding* ini. Berikut adalah hasil pembelajaran XGBClassifier:



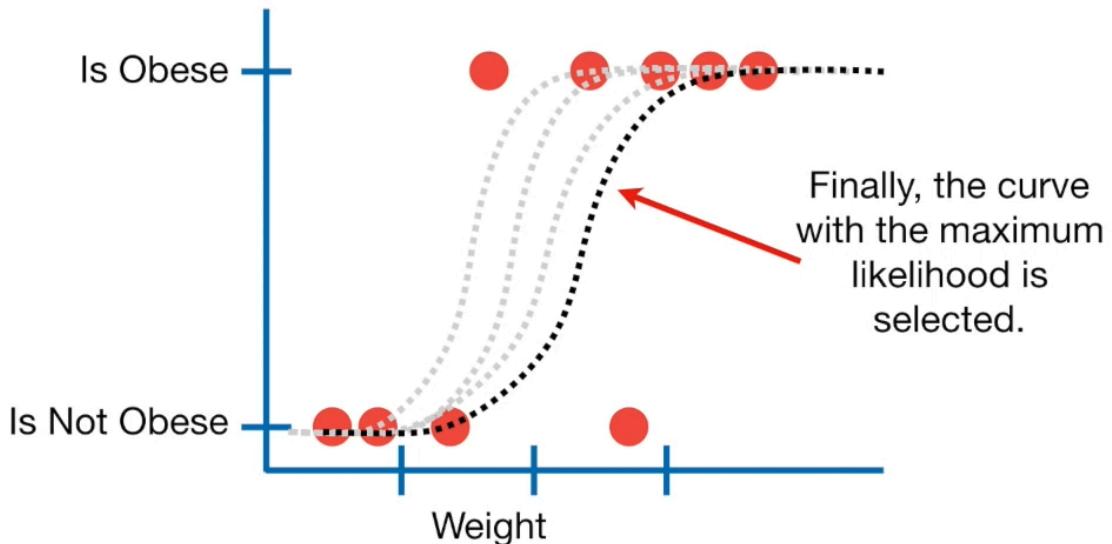
Berikut adalah interpretasi objektif yang penting dari ilustrasi diatas:

- 3 prediktor terkuat untuk memprediksi kolesterol total adalah trigliserida, usia, dan lemak ('Fat').
- Trigliserida adalah prediktor yang paling kuat. Kolesterol total dan trigliserida memiliki korelasi positif.
- Perempuan berusia tua rentan memiliki kolesterol tinggi.
- Semakin lama masa kerja seseorang, semakin tinggi total kolesterol.
- IMT memiliki korelasi positif dengan kolesterol total.
- Tekanan darah sistol ('Tekanan darah (S)') dan tekanan darah diastol ('Tekanan darah (D)') berkorelasi positif dengan kolesterol total.

Logistic Regression

Logistic Regression adalah metode statistik untuk menganalisis hubungan variabel-variabel, dan hasil prediksi biner (Ranganathan et al, 2017). Metode ini cocok untuk masalah yang diberikan, yaitu memprediksi apakah kolesterol seseorang tinggi (*True/1*) atau rendah (*False/0*).

Logistic Regression bekerja dengan mencocokkan fungsi *sigmoid* (Mutea, n.d.,), atau fungsi yang berbentuk 'S' yang digunakan untuk prediksi biner. Kemudian, dilakukan sebuah algoritma yang dinamakan *maximum likelihood*. Algoritma tersebut bekerja dengan menghitung probabilitas benar atau salah setiap barisan data. Kemudian, semua probabilitas dikali. Garis fungsi *sigmoid* akan digeser berkali-kali (mengubah kecondongan bentuk 'S') hingga mendapatkan hasil kali probabilitas yang terbesar



(Sumber grafik: “StatQuest: Logistic Regression,” <https://www.youtube.com/watch?v=yIYKR4sgzI8>)

Akan tetapi, tidak banyak yang dipelajari melalui model ini, yang akan ditunjukkan di bagian “Hasil Prediksi.” Ada 2 alasan yang bisa menjelaskan ini:

- *Encoding* kolom ‘Cholesterol Total (mg/dL)’ menjadi 0 dan 1 menghilangkan banyak informasi penting, yang hanya bisa didapat dengan membiarkan nilai-nilai tersebut sebagai angka.
- Karena *Logistic Regression* merupakan model regresi linier, maka prediksi yang melibatkan proses tidak linier tidak dapat dilakukan secara akurat.
- *Logistic Regression* memiliki kelemahan dalam menganalisis relasi antar fitur-fitur yang dimasukkan kedalamnya.

Hasil Prediksi

Pada bagian ini, akan ditunjukkan nilai akurasi berdasarkan metrik yang relevan untuk model yang dibuat, dan juga ilustrasi pembelajaran model. Adapun beberapa metrik dan ilustrasi yang digunakan untuk menganalisis kualitas model:

- RMSE (*Root Mean Square Error*): metrik ini mengukur rata-rata selisih antara nilai yang diprediksi, dengan nilai sesungguhnya (“Root Mean Squared Error (RMSE)”, n.d.,). Semakin kecil RMSE, semakin

$$RMSE = \sqrt{\frac{SSE_W}{W}} = \sqrt{\frac{1}{W} \sum_{i=1}^N w_i u_i^2}$$

bagus sebuah model.

- *Confusion Matrix*: ilustrasi yang menggambarkan kemungkinan hasil prediksi (“*What Is a Confusion Matrix in Machine Learning?*”, 2024). Pada kasus ini, hasil prediksi dipetakan berdasarkan kesesuaian antara nilai *boolean* prediksi dengan nilai *boolean* sesungguhnya.

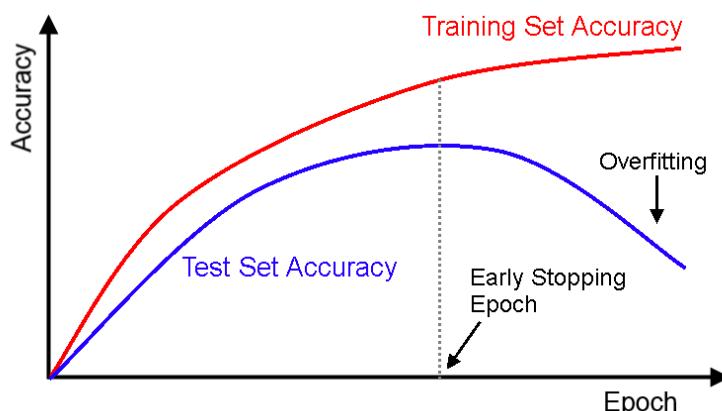
		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

- TP (*True Positive*, atau *precision*) = nilai sungguhan positif (*true* atau 1), prediksi positif.
- FP (*False Positive*, atau *recall*) = nilai sungguhan negatif (*false* atau 0), prediksi positif.
- FN (*False Negative*) = nilai sungguhan positif, prediksi negatif.
- TN (*True Negative*) = nilai sungguhan negatif, prediksi negatif.
 - Idealnya, model harus memiliki *precision* dan *recall* yang tinggi untuk prediksi yang lebih aman ketika diaplikasikan pada sebuah masalah.

- *F1 Score*: metrik ini dikalkulasikan dengan menemukan nilai rata-rata harmonik antara *precision* dan *recall*. Nilai terbaik F1 adalah 1, dan nilai terburuk adalah 0 (“*Sklearn.Metrics.F1_score.*”, n.d.,).

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}$$

- *Learning Curve*: ilustrasi ini menunjukkan seberapa cepat sebuah model *machine learning* menjadi lebih baik seiring jumlah data *training* bertambah (Amari et al, 1993). Idealnya, model bisa menambah akurasi atau mengurangi error selama *training*, dan bisa men-generalisasi pola pada *training* pada data *test* ataupun data baru (Ibrahim, 2023).



- *Learning Graph*: grafik yang menunjukkan perbandingan nilai prediksi dengan nilai sesungguhnya. Idealnya, model harus bisa menyesuaikan dengan garis yang menggambarkan hubungan antara kedua nilai. Sumbu X adalah nilai sungguhan, dan sumbu Y adalah nilai yang diprediksi model.
- *Feature Importance*: ilustrasi ini menggambarkan seberapa penting sebuah fitur, variabel atau kolom, dalam menjelaskan sifat model (Saarela et al, 2021). Dalam kasus ini, ilustrasi ini digunakan untuk memilih variabel yang pantas digunakan untuk prediksi. Fitur diurutkan dari atas, menurut skala kepentingan.

Untuk beberapa model, akurasi diukur melalui metode *cross-validation* (CV), atau melatih model pada setiap bagian data *training*. Ini dilakukan agar model tidak *overfit*, atau terpaku terhadap porsi data tes saja (Cook, n.d.). Maka dari itu, beberapa model akan diukur kualitasnya menggunakan *learning curve*.

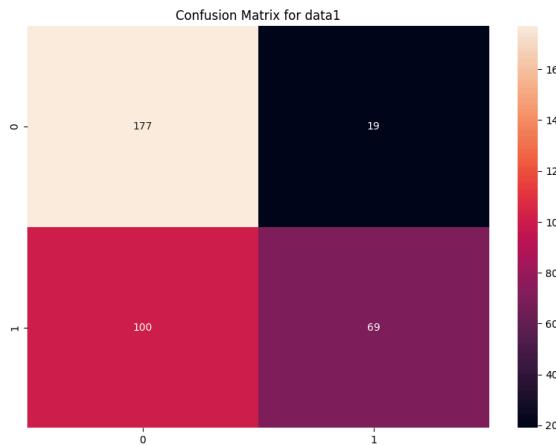
Satu catatan penting adalah, untuk prediksi regresi, data-data yang tidak seimbang ('Cholesterol Total (mg/dL) = 187.0) dibuang, dan disisakan 20 data dengan nilai tersebut.

```
Cholesterol Total (mg/dL)
187.0    790
174.0     14
180.0     12
206.0     11
203.0     11
...
229.0      1
127.0      1
146.0      1
308.0      1
249.0      1
Name: count, Length: 144, dtype: int64
```

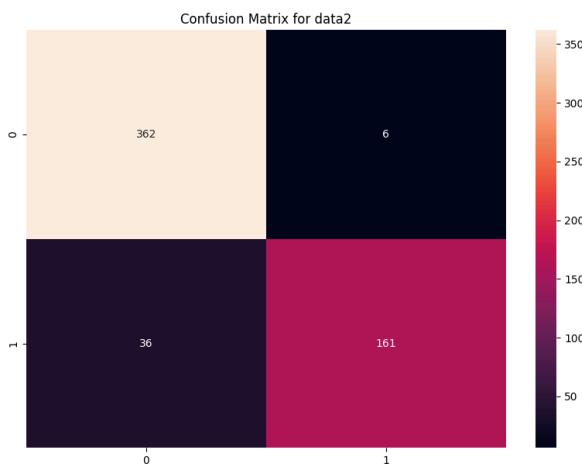
Ini dilakukan agar model regresi tidak *overfit* terhadap satu nilai saja. Akan tetapi, ini tidak dilakukan untuk model klasifikasi, karena model yang digunakan dapat mengatasi ketidakseimbangan tersebut dengan lebih baik. Maka dari itu, data 2 di pisah menjadi data *training* dan data *test*, agar pelatihan model dan pengujian model menjadi lebih baik. Setelah itu, kualitas model diuji dengan data 1, yaitu data baru yang tidak ada pada *training* maupun *test*.

XGBRegressor:

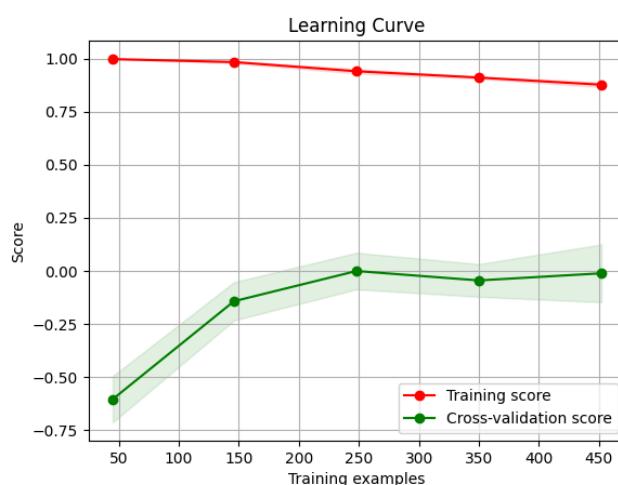
- Akurasi tes dari CV (*cross-validation*) scoring: 31.966111547842956
- RMSE data2: 11.879291034227295
- RMSE data1: 30.452187269299262
- *F1 score* dari model terbaik pada data2: 0.8846153846153846
- *F1 score* dari model terbaik pada data1: 0.5369649805447471
- *Confusion Matrix* terhadap data 1:



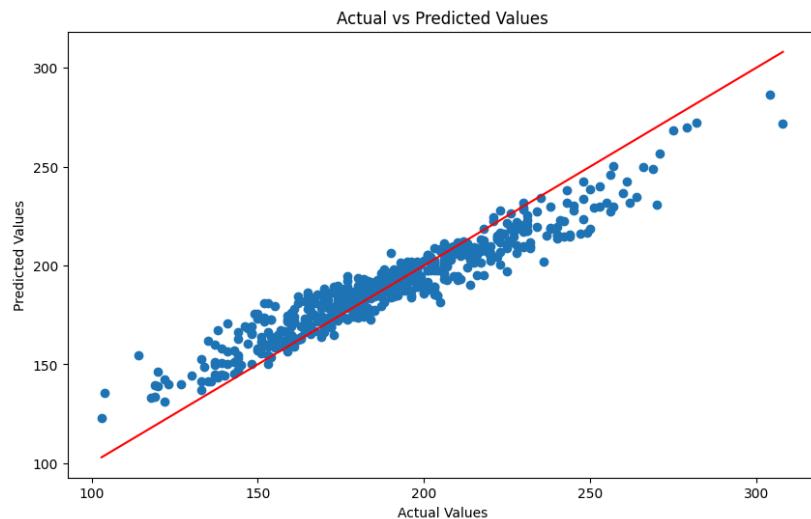
- *Confusion Matrix* terhadap data 2:



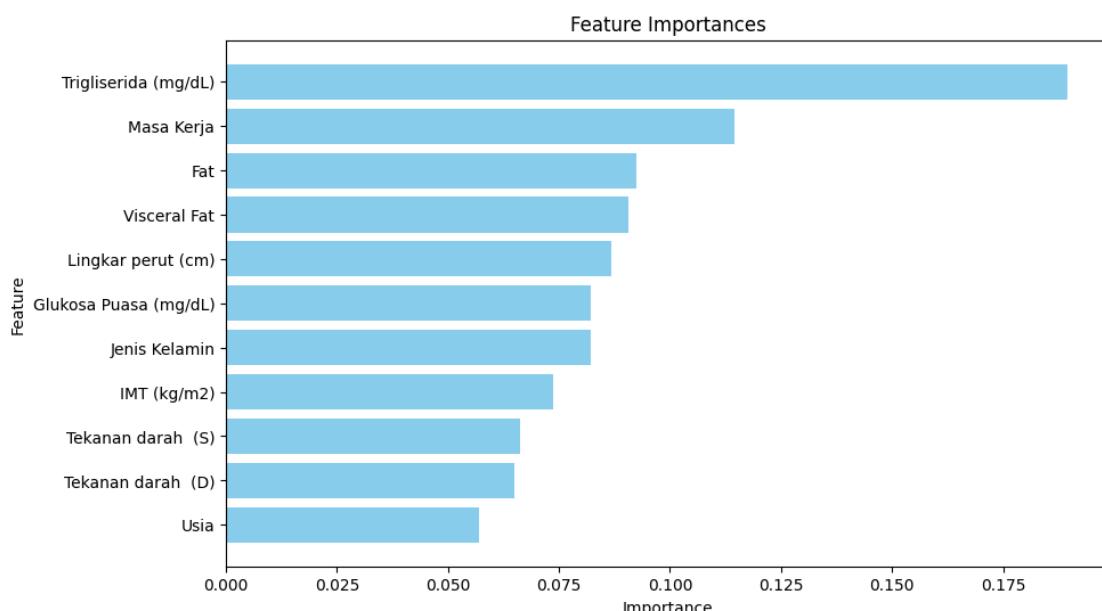
- *Learning Curve*:



- Grafik Sungguhan vs Prediksi:



- *Feature Importance*



XGBClassifier:

- *Confusion matrix* untuk data *training*

672	229
0	162

```
Recall on train set: 1.0
Precision on train set: 0.4143222506393862
F1 on train set: 0.5858951175406871
```

- *Confusion matrix* untuk data *test*:

173	59
1	33

```
Recall on test set: 0.9705882352941176
Precision on test set: 0.358695652173913
F1 on test set: 0.523809523809523
```

- *Confusion matrix* untuk data 1:

43	153
11	158

```
Recall on data 1: 0.9349112426035503
Precision on data 1: 0.5080385852090032
F1 on data 1: 0.6583333333333333
```

Logistic Regression:

- *Confusion matrix* untuk data *training*:

672	229
0	162

```
Recall on train set: 1.0
Precision on train set: 0.4143222506393862
F1 on train set: 0.5858951175406871
```

- Confusion matrix untuk data *test*:

173	59
1	33

```
Recall on test set: 0.9705882352941176
Precision on test set: 0.358695652173913
F1 on test set: 0.5238095238095238
```

- Confusion matrix untuk data 1:

43	153
11	158

```
Recall on data 1: 0.9349112426035503
Precision on data 1: 0.5080385852090032
F1 on data 1: 0.6583333333333333
```

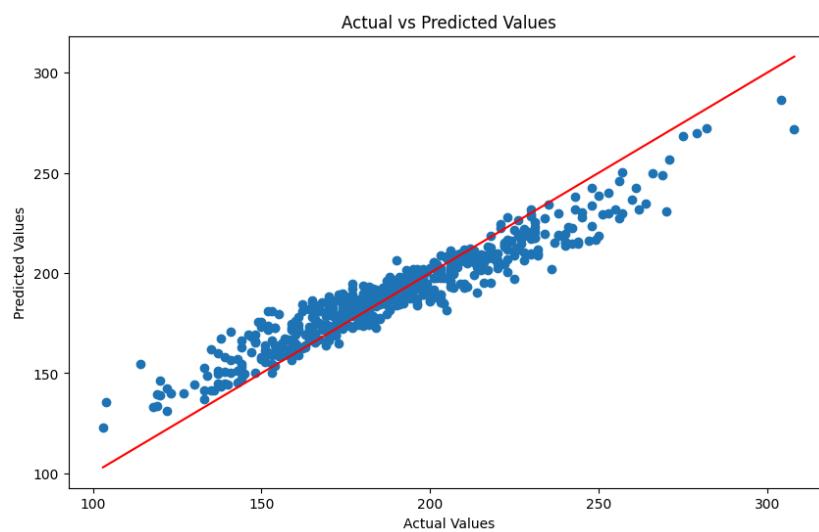
Bab III Penutup

Kesimpulan dan Saran

Setelah melakukan analisis matematis menggunakan *machine learning*, dapat disimpulkan bahwa 3 faktor yang paling berpengaruh pada kadar total kolesterol seseorang adalah sebagai berikut:

1. Triglicerida
2. Masa kerja
3. (kadar) lemak

Untuk melakukan prediksi akurat kolesterol total seseorang, pendekatan regresi menggunakan XGBRegressor adalah pendekatan yang lebih baik. Ini didukung dengan learning graph yang menunjukkan bahwa prediksi model menyerupai nilai kolesterol total sesungguhnya.



Akan tetapi, untuk aplikasi medis, lebih disarankan untuk menggunakan XGBClassifier, dikarenakan jumlah *recall* yang tinggi. Asumsi ini akan lebih aman untuk membuat keputusan tentang kondisi kesehatan seseorang. Sebagai contoh, model lebih baik berasumsi bahwa seseorang memiliki kolesterol ketika sebetulnya tidak ada (*False Positive, recall*), daripada berasumsi bahwa seseorang tidak memiliki kolesterol ketika sebetulnya ada.

43	153
11	158

- TP (*True Positive*, atau *precision*) = 43
- FP (*False Positive*, atau *recall*) = 153
- FN (*False Negative*) = 11
- TN (*True Negative*) = 158

Here's your reference list formatted in APA style and ordered alphabetically:

Daftar Pustaka

- Aguinis, H., Gottfredson, R., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods, 16*, 270-301.
<https://doi.org/10.1177/1094428112470848>
- Amari, S. (1993). A universal theorem on learning curves. *Neural Networks, 6*, 161-166.
[https://doi.org/10.1016/0893-6080\(93\)90013-M](https://doi.org/10.1016/0893-6080(93)90013-M)
- Andrade, C. (2019). The p Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives. *Indian Journal of Psychological Medicine*.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6532382/>
- Behrens, J. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*, 131-160. <https://doi.org/10.1037/1082-989X.2.2.131>

Brownlee, J. (2020). How to Tune the Number and Size of Decision Trees with XGBoost in Python.
<https://machinelearningmastery.com/tune-number-size-decision-trees-xgboost-python/>

Centers for Disease Control and Prevention. (2022, June 3). Assessing your weight.
www.cdc.gov/healthyweight/assessing/index.html

Chatterjee, S., & Hadi, A. (1986). Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science, 1*, 379-393. <https://doi.org/10.1214/SS/1177013622>

Chip. (2024, January 11). How to Read a Correlation Heatmap.
www.quanthub.com/how-to-read-a-correlation-heatmap/

Chu, X., & Ilyas, I. (2016). Qualitative Data Cleaning. *Proc. VLDB Endow., 9*, 1605-1608.
<https://doi.org/10.14778/3007263.3007320>

Cook, A., & DanB. (2023). Scaling and Normalization.
www.kaggle.com/code/alexisbcook/scaling-and-normalization

Cook, A., & DanB. (2023). XGBoost. www.kaggle.com/code/alexisbcook/xgboost

Cook, A., & DanB. (2023). Cross-Validation. www.kaggle.com/code/alexisbcook/cross-validation

Emerson, J., Green, W., Schloerke, B., Crowley, J., Cook, D., Hofmann, H., & Wickham, H. (2013). The Generalized Pairs Plot. *Journal of Computational and Graphical Statistics, 22*, 79-91. <https://doi.org/10.1080/10618600.2012.694762>

Freedman, D., Gruchow, H., Anderson, A., Rimm, A., & Barboriak, J. (1988). Relation of triglyceride levels to coronary artery disease: the Milwaukee Cardiovascular Data Registry. *American Journal of Epidemiology, 127*(6), 1118-1130. <https://doi.org/10.1097/00008483-198810000-00016>

Goswami, S. (2020, November 2). How to use SMOTE, Borderline SMOTE, ADASYN to handle class imbalance. [Video]. www.youtube.com/watch?v=mKG7lnZNAOk

Gu, Z., & Hübschmann, D. (2021). Make Interactive Complex Heatmaps in R. *F1000Research*. <https://doi.org/10.7490/F1000RESEARCH.1118617.1>

Howe, D. (2001). Redundant vs duplicated data. <https://doi.org/10.1016/B978-075065086-1/50005-8>

Ibrahim, M. (2023). A Deep Dive Into Learning Curves in Machine Learning. wandb.ai/mostafaibrahim17/ml-articles/reports/A-Deep-Dive-Into-Learning-Curves-in-Machine-Learning--Vmlldzo0NjA1ODY0

Kathula, A. (2023). How to Handle Imbalanced Data in Classification. www.phdata.io/blog/how-to-handle-imbalanced-data-in-classification/

Khurana, U., Samulowitz, H., & Turaga, D. (2017). Feature Engineering for Predictive Modeling using Reinforcement Learning. *ArXiv, abs/1709.07150*. <https://doi.org/10.1609/aaai.v32i1.11678>

Krishnan, S., Wang, J., Wu, E., Franklin, M., & Goldberg, K. (2016). ActiveClean: Interactive Data Cleaning While Learning Convex Loss Models. *ArXiv, abs/1601.03797*

Lev, A. (2022). XGBoost versus Random Forest. www.qwak.com/post/xgboost-versus-random-forest

Low, W., Lee, M., & Ling, T. (2001). A knowledge-based approach for duplicate elimination in data cleaning. *Inf. Syst., 26*, 585-606. [https://doi.org/10.1016/S0306-4379\(01\)00041-2](https://doi.org/10.1016/S0306-4379(01)00041-2)

Morgenthaler, S. (2018). Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 1*. <https://doi.org/10.1002/wics.2>

Morrison, D. (2010). Using data-display networks for exploratory data analysis in phylogenetic studies. *Molecular Biology and Evolution, 27*(5), 1044-1057. <https://doi.org/10.1093/molbev/msp309>

Mutea, B. Logistic Regression in Python with Scikit-Learn.
<https://www.machinelearningnuggets.com/logistic-regression/>

National Heart Lung and Blood Institute. (n.d.). Causes and Risk Factors.
www.nhlbi.nih.gov/health/blood-cholesterol/causes

Penn Medicine. (n.d.). High cholesterol.
www.pennmedicine.org/for-patients-and-visitors/patient-information/conditions-treated-a-to-z/high-cholesterol

Ranganathan, P., Pramesh, C., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in Clinical Research, 8*, 148-151. https://doi.org/10.4103/picr.PICR_87_17

Rochimawati. (2016, November 3). Dua kawasan ini akan jadi kota metropolitan baru. www.viva.co.id/arsip/842969-dua-kawasan-ini-akan-jadi-kota-metropolitan