Times New Roman

Prediksi *Total Fat* dengan Machine Learning

Bogo Bogo Sort

Darrel Danadyaksa Poli

Patrick Samuel Evans Simanjuntak

Yesaya Rudolf

16 Maret 2024

Pendahuluan

A. Latar Belakang. Kesehatan masyarakat merupakan isu yang mendesak dan terus memerlukan inovasi dalam pendekatan dan solusi. Berbagai tindakan dilakukan untuk meningkatkan kesehatan baik dari segi preventif maupun represif. Berbagai hal seperti vaksinasi, makna makanan bergizi, dan konsumsi vitamin serta mineral yang cukup merupakan hal preventif yang dapat dilakukan. Di sisi lain, tindakan represif seperti obat merupakan standard yang dilakukan untuk mengobati pasien yang sakit. Akan tetapi, pemberian obat hanya boleh diberikan atas diagnosis dari dokter. Diagnosis ini didasari oleh beberapa hal. Salah satunya adalah tes darah. Tes darah merupakan suatu acuan untuk mengecek kondisi tubuh dari seseorang yang dicurigai mengidap suatu penyakit. Meskipun tes darah telah menjadi standar untuk diagnosis dan pemantauan kondisi kesehatan, masih ada tantangan dalam hal aksesibilitas, biaya, dan kepraktisan. Faktor-faktor ini mempengaruhi berbagai kalangan, terutama mereka yang tinggal di daerah terpencil atau memiliki kondisi medis yang mempersulit proses pengambilan sampel darah.

Teknologi *machine learning* menjanjikan kemungkinan untuk mengatasi beberapa tantangan ini. Dengan kemampuannya untuk menganalisis data kompleks dan menemukan pola-pola yang tersembunyi, *machine learning* dapat memberikan wawasan baru dalam pemantauan kesehatan dan prediksi penyakit.

B. Rumusan Masalah. Selain tantangan aksesibilitas dan biaya, tes darah juga memiliki keterbatasan dalam hal frekuensi. Seringkali, tes darah hanya dilakukan pada titik waktu tertentu, yang mungkin tidak mencerminkan kondisi kesehatan secara menyeluruh. Misalnya, seseorang dapat memiliki kadar kolesterol yang tinggi hanya pada beberapa saat tertentu, yang mungkin tidak terdeteksi dalam tes darah yang dilakukan pada waktu lain.

Oleh karena itu, ada kebutuhan untuk memperkenalkan metode yang dapat memberikan pemantauan kontinu dan prediksi dinamis terkait dengan kadar kolesterol. Penerapan *machine learning* dalam hal ini dapat membantu dalam mengatasi keterbatasan tersebut.

C. Tujuan. Tujuan utama dari penelitian ini adalah untuk mengembangkan model prediksi kadar kolesterol yang dapat bekerja secara kontinu dan memberikan perkiraan yang akurat tanpa perlu pengambilan sampel darah yang konvensional. Selain itu, kami bertujuan untuk mengeksplorasi berbagai faktor lain yang dapat memengaruhi kadar kolesterol, seperti pola makan, aktivitas fisik, dan faktor genetik.

Dengan demikian, penelitian ini tidak hanya bertujuan untuk menciptakan alternatif bagi tes darah konvensional, tetapi juga untuk memberikan pemahaman yang lebih dalam tentang faktor-faktor yang memengaruhi kadar kolesterol. Hal ini diharapkan dapat memberikan dasar yang kuat untuk pengembangan strategi pencegahan dan manajemen yang lebih efektif dalam mengurangi risiko penyakit kardiovaskular dan meningkatkan kualitas hidup secara keseluruhan.

Dalam artikel ini, kami akan membahas secara rinci tentang metodologi yang kami gunakan, analisis hasil, serta implikasi temuan kami dalam konteks kesehatan masyarakat dan perkembangan ilmu pengetahuan kedokteran secara luas. Selain itu, kami juga akan menyoroti tantangan dan peluang yang mungkin muncul dalam penerapan model prediksi ini dalam praktik klinis dan penelitian lanjutan.

Pembahasan

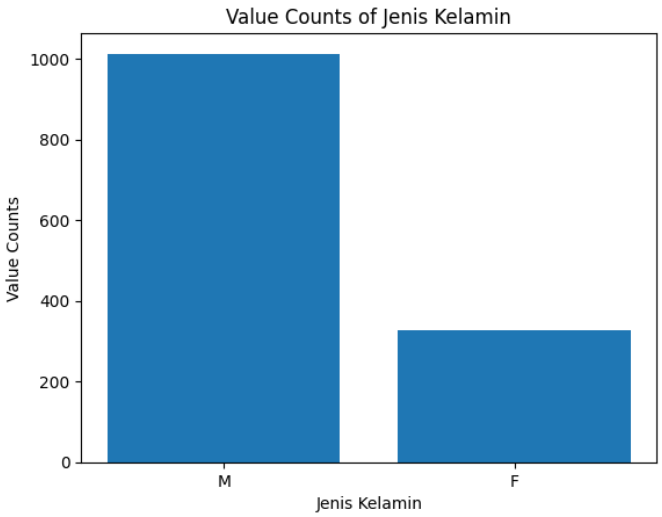
D. *Explanatory Data Analysis* (EDA). Langkah pertama dalam implementasi adalah *data cleaning*. Data ini tidak memiliki *missing values sama sekali*. Pada proses Data Cleaning, kita perlu melihat kondisi data dan persebaran dari masing-masing variabel yang ada pada data.

1. Responden

Variabel ini merupakan variabel yang serupa dengan id dimana setiap angka unik digunakan untuk merepresentasikan suatu data pada *dataset*. Variabel ini nantinya akan di-*drop* dari *dataset* karena tidak merepresentasikan apapun.

2. Jenis Kelamin

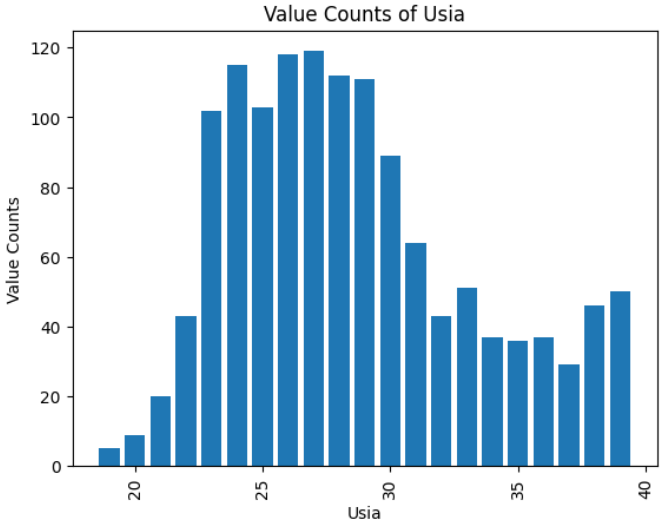
Jenis kelamin adalah variabel yang memiliki dua nilai unik, yaitu laki-laki dan perempuan. Berikut adalah persebaran dari variabel jenis kelamin.



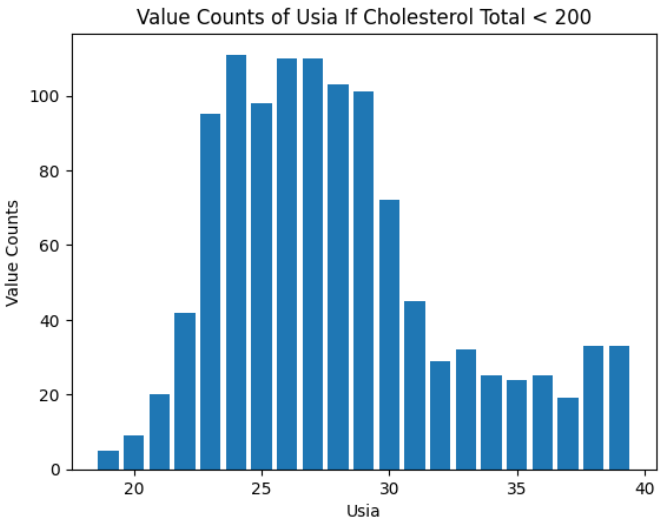
Dapat diamati bahwa jenis kelamin dari sampel yang ada didominasi oleh laki-laki dengan 1013 data dan diikuti oleh perempuan dengan 326 data. Dapat disimpulkan bahwa data ini memiliki bias pada jenis kelamin laki-laki karena adanya *imbalance* pada variabel ini. Variabel ini tidak memerlukan *cleaning* apapun karena jenis datanya yang biner dan tidak memiliki *missing values* sama sekali.

3. Usia

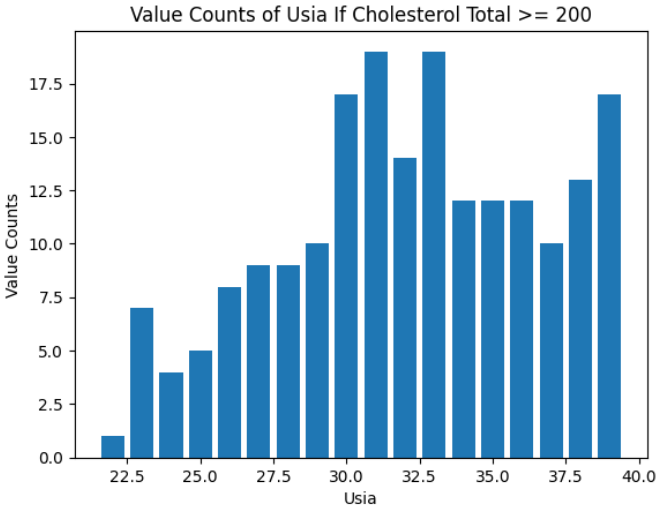
Variabel usia adalah variabel yang merepresentasikan usia dari masing-masing sampel data. Berikut adalah persebaran dari variabel usia.



Dari *bar plot* tersebut. dapat dilihat bahwa mayoritas data merupakan individu dengan usia 25-30 tahun. Data juga hanya mencakup usia 19 hingga 39 tahun. Artinya, model yang akan dibuat memiliki bias pada kisaran usia tersebut. Kita akan melihat hubungan variabel ini dengan kolom target, yaitu kolesterol. Pertama kita akan melihat persebaran usia dari sampel dengan kadar kolesterol total dibawah 200 mg/dL (kadar kolesterol normal).



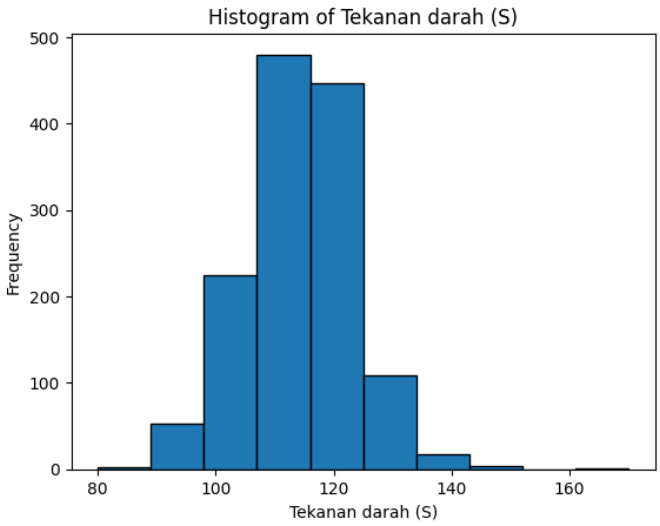
Dapat dilihat bahwa persebarannya tersebar pada usia 23-30 tahun. Hal ini berarti kebanyakan orang yang memiliki kadar kolesterol yang normal adalah orang dengan usia dibawah 30 tahun. Dibawah ini kita dapat melihat persebaran usia dari sampel dengan kadar kolesterol total diatas atau sama dengan 200 mg/dL (kadar kolesterol tinggi).



Dapat dilihat bahwa rentang usia 23-30 pada sampel dengan kolesterol >= 200 mg/dL memiliki jumlah yang lebih sedikit dibandingkan dengan sampel dengan kolesterol < 200 mg/dL. Variabel ini pun tidak memerlukan *data cleaning* sama sekali.

4. Tekanan darah (S)

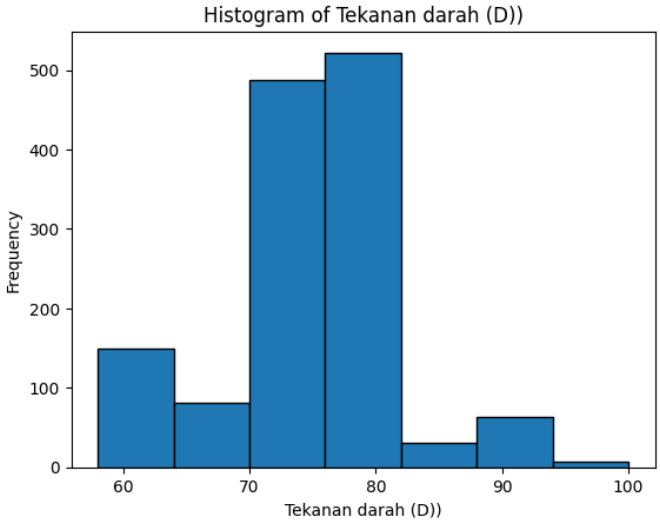
Tekanan darah (S) merupakan variabel yang merepresentasikan tekanan darah sistolik, yaitu tekanan darah ketika jantung memompa darah ke seluruh tubuh. Berikut adalah persebaran datanya.



Dari histogram tersebut, terlihat bahwa persebaran dari data tersebut tidak normal.

5. Tekanan darah (D)

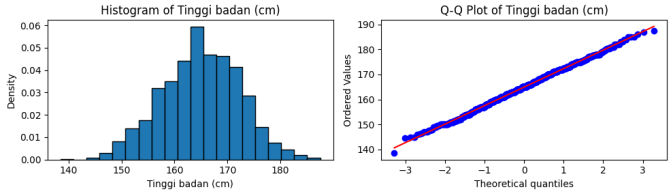
Tekanan darah (D) merupakan variabel yang merepresentasikan tekanan darah diastolik, yaitu tekanan darah ketika jantung mengalami relaksasi (tidak memompa). Berikut adalah persebaran datanya.



Dari histogram tersebut, terlihat bahwa persebaran dari data tersebut tidak normal.

6. Tinggi badan

Pertama perlu diperhatikan bahwa tinggi badan merupakan sebuah fitur yang menarik yang lain dari fitur lainnya. Hal ini dapat dikatakan karena karakteristik dari fitur kolom ini yang mengikuti distribusi normal. Hal ini dapat kita lihat dari inspeksi visual.

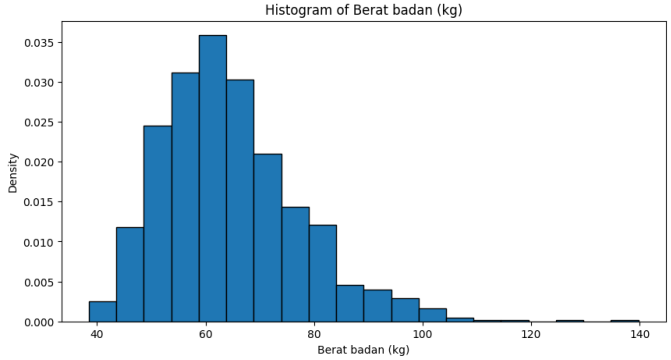


Selain itu, dengan uji statistik, kita juga bisa melihat bahwa skor *p-value* dari uji *Shapiro-Wilk* adalah $9.218852222197998 \cdot 10^{-18}$. Sedangkan, skor textitp-value dari uji *D'Agostino and Pearson's* adalah $5.588443686434393 \cdot 10^{-36}$

Kedua inspeksi ini dapat membuat kita dapat menyimpulkan bahwa fitur ini terdistribusi secara merata.

7. Berat Badan

Perhatikan bahwa fitur ini secara keseluruhan tidak memiliki perubahan signifiikan pada seluruh data yang ada, oleh karena itu, fitur ini tidak akan diubah atau dimodifikasi. Berikut adalah tampilan *plot* distribusi dari fitur ini.



8. IMT

Perlu diperhatikan bahwa kolom ini adalah kolom yang menarik. Dari analisis yang telah didapatkan, didapati bahwa kolom ini merupakan hasil dari kolom berat badan yang dibagi dengan kuardrat dari tinggi badan. Dari operasi tersebut, didapati bahwa selisihnya sangat sedikit dengan kolom IMT yang telah diberikan. Oleh karena itu, kami memutuskan untuk menghapus kolom IMT dan tidak menggunakannya, karena kita sudah memiliki informasi berat badan dan tinggi badan yang sesungguhnya.

9. Lingkar Perut

Dengan melakukan inspeksi visual dan *statistical test* yang serupa pada kolom tinggi badan, dapat disimpulkan bahwa fitur ini adalah fitur yang terdistribusi secara normal. Oleh karena itu kami tidak mengubah fitur ini.

10. Glukosa Puasa dan trigliserida

Perhatikan bahwa melaui *plotting* yang dilakukan, dapat dilihat bahwa distribusi data dari kolom-kolom ini mirip. Persamaan dari kedua kolom ini adalah bagaimana kita da-

pat melihat bahwa persebaran datanya memiliki sebuah nilai yang jauh lebih banyak kemunculannya daripada data yang lainnya. Kemudian, tidak ada treatment yang bisa dilakukan untuk data-data ini karena tidak ada perbedaan efek dari kedua kolom ini terhadap hal yang ingin kita perdiksi. Tetapi pada orang yang memiliki kolesterol total sama dengan 187 mg/dL, terdapat 783 orang yang memiliki Trigliserida sebesar 99 mg/dL.

Value Counts Trigliserida pada Orang dengan Kolesterol Total 187 md/dL

Trigliserida (mg/dL)	Value Counts
99.0	783
78.0	1
47.0	1
88.0	1
76.0	1
634.0	1
84.0	1
129.0	1

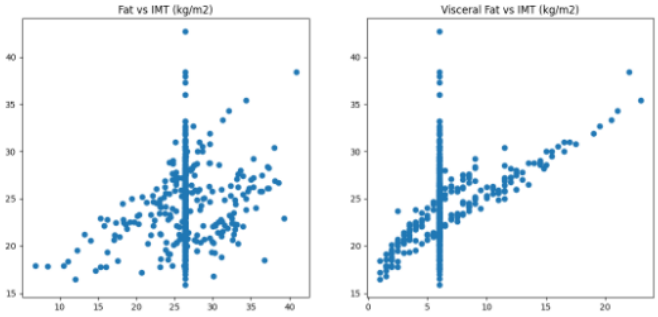
Value Counts Trigliserida pada Orang dengan Kolesterol Total Bukan 187 md/dL

Trigliserida (mg/dL)	Value Counts
80.0	10
81.0	9
82.0	9
68.0	8
77.0	8
..	..
219.0	1
179.0	1
210.0	1
203.0	1
331.0	1

Maka fitur Trigliserida dapat digunakan sebagai *classifier* untuk menentukan apakah kolesterol total merupakan 187 mg/dL atau bukan.

11. Fat dan Visceral Fat

Perhatikan bahwa kedua kolom ini memiliki persebaran data yang menarik juga karena jika Kolesterol totalnya bukan 187, maka persebaran data dari Fat dan Visceral fat akan memiliki data yang memiliki jumlah kemunculan yang lebih sering jika dibandingkan dengan data lainnnya. Untuk memperbaiki hal ini, akan dilakukan koreksi dari data yang ada dengan mengimplementasikan regresi linear biasa untuk nilai yang kemunculannya sering yang dapat dilihat korelasinya dengan fitur IMT (kg/m2) seperti yang terlihat pada plot berikut ini.



Value Counts Fat pada Orang dengan Kolesterol Total Bukan 187 md/dL

Fat	Value Counts
26.4	339
25.2	5
27.9	5
29.6	5
23.1	4
..	..
18.8	1
25.5	1
26.8	1
34.6	1
17.1	1

Value Counts Visceral Fat pada Orang dengan Kolesterol Total Bukan 187 md/dL

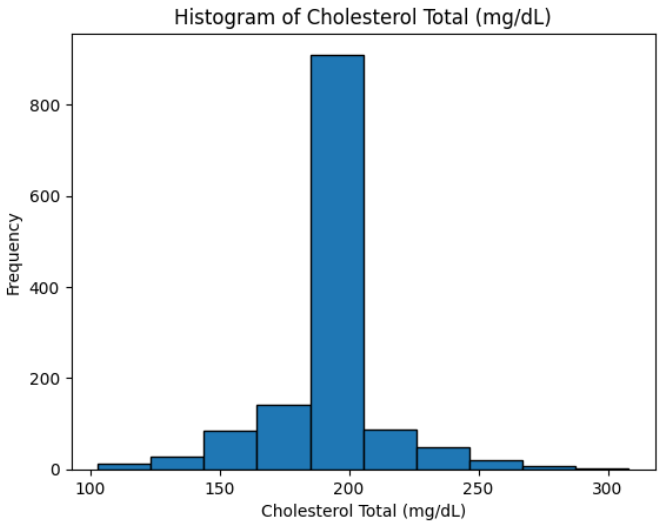
Visceral Fat	Value Counts
6.0	354
2.5	12
2.0	11
4.0	10
8.0	10
..	..
21.0	1
23.0	1
16.0	1
17.0	1

12. Tempat Lahir

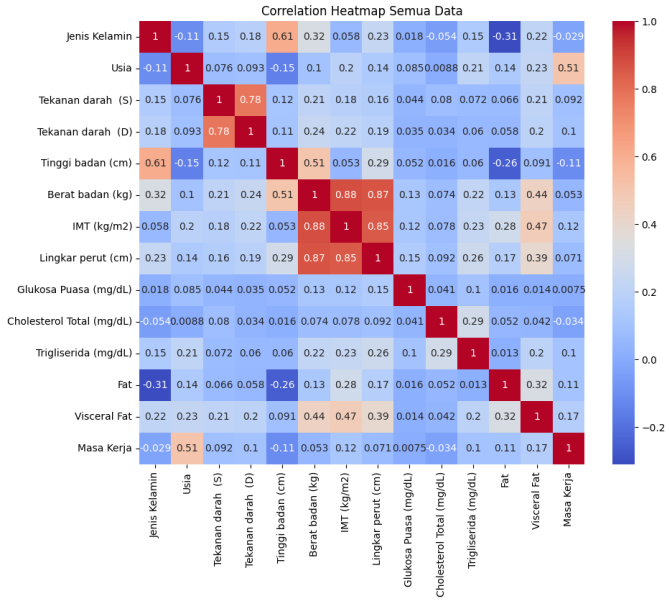
Perlu diperhatikan bahwa tempat lahir tidaklah dapat menentukan kadar kolesterol secara langsung, dikarenakan faktor kesehatan lebih dipengaruhi oleh genetika atau pola hidup.

13. Cholesterol Total (mg/dL)

Variabel ini merupakan variabel yang akan diprediksi oleh model. Variabel ini memiliki persebaran yang hanya terfokus pada satu titik, yaitu 187.

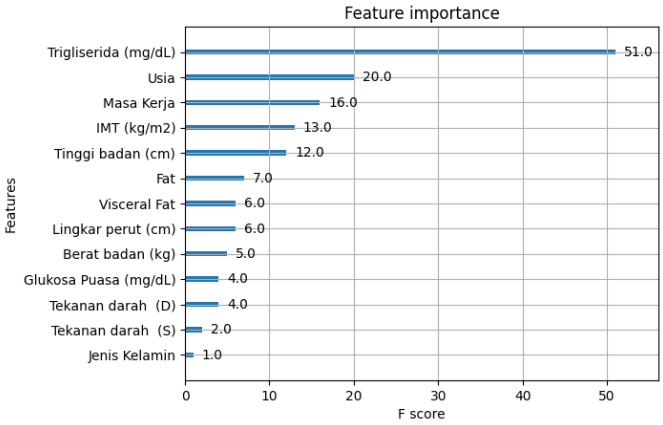


Untuk korelasi antar variabel dapat dilihat pada *correlation heatmap* sebagai berikut

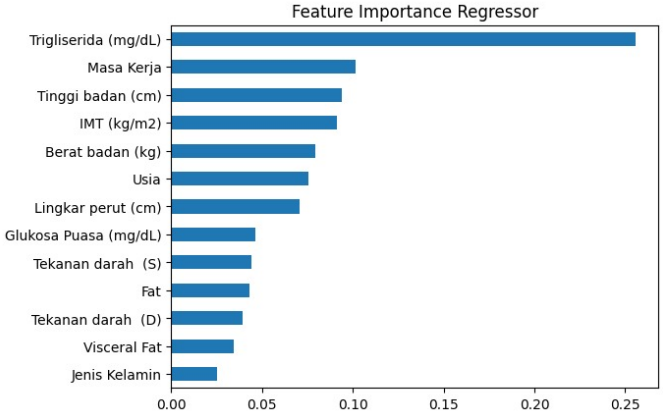


14. Fitur yang Paling Berpengaruh dalam Prediksi

Dalam menentukan fitur yang paling berpengaruh dalam prediksi, kami mencoba dua metode, yaitu feature importance dari XGBoost dan Chi Squared Test. - Feature Importance dari XGBoost Peneliti menggunakan XGBoost dengan importance berupa "gain" dimana *feature importance* dihitung dari seberapa sering sebuah fitur muncul pada *tree* yang ada pada model. Dalam menentukan *conditional* fitur dari suatu node, model akan melakukan perhitungan untuk mencari *splitting point* yang akan menghasilkan gain terbesar. Berikut adalah *feature importance* dari *classifier* yang menentukan apakah nilai dari total fat seseorang adalah 187.

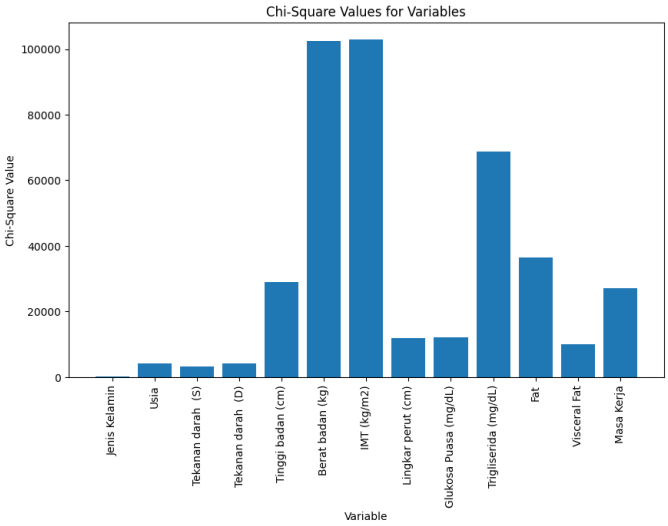


Selanjutnya, berikut adalah *feature importance* dari masing-masing variabel dari model Random Forest *regressor* untuk data dengan total fat yang tidak sama dengan 187.

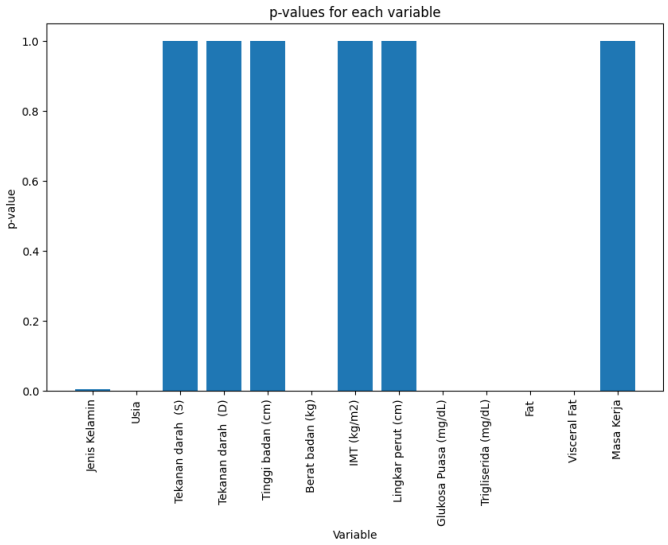


Terlihat bahwa variabel Trigelisterida selalu menempati posisi sebagai fitur yang paling penting. Hal ini sebenarnya disebabkan karena rumus dari Total Fat adalah high-density lipoprotein (HDL) cholesterol + low-density lipoprotein (LDL) cholesterol + 20% triglycerides, sehingga kadar trigelisterida mempengaruhi Total Fat secara langsung. Selain trigelisterida, terdapat variabel-variabel lain yang penting, seperti Usia, Masa Kerja, IMT, Tinggi Badan, Berat Badan. Variabel-variabel lain memiliki pengaruh yang insignifikan dalam mempengaruhi model.

Kedua, kita akan mencoba menggunakan chi square test untuk menentukan apakah suatu data memiliki nilai chi squared test yang cukup besar mengartikan bahwa korelasi antara suatu variabel dan variabel target cukup besar. Berikut adalah visualisasi dari nilai chi square dari masing-masing variabel terhadap variabel target.



Berikut adalah p-value dari masing-masing variabel target. Kita akan menggunakan significance level=0.95 dimana variabel dengan p-value>0.05 akan tidak dipakai karena gagal menolak H0 dimana H0 adalah variabel tidak memiliki hubungan signifikan dengan variabel target. Sebaliknya, p-value<0.05 berhasil menolak H0. Berikut adalah visualisasinya.

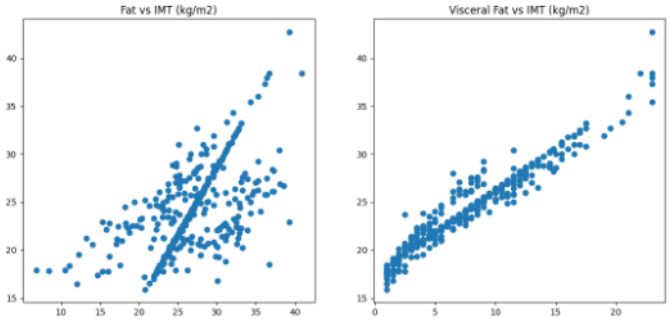


E. *Data Cleaning*. 1. Fat dan Visceral Fat Dapat dilakukan pemisahan data Fat antara 26.4 atau bukan dan pemisahan data Visceral Fat antara 6 atau bukan dikarenakan memiliki persebaran yang tidak merata. Setelah dilakukan pemisahan data yang bukan merupakan nilai mayoritas, dilakukanlah regresi linear dan didapatkan model sebagai berikut.

$$\hat{y_{\text{Fat}}} = 0.706168x + 9.747593$$

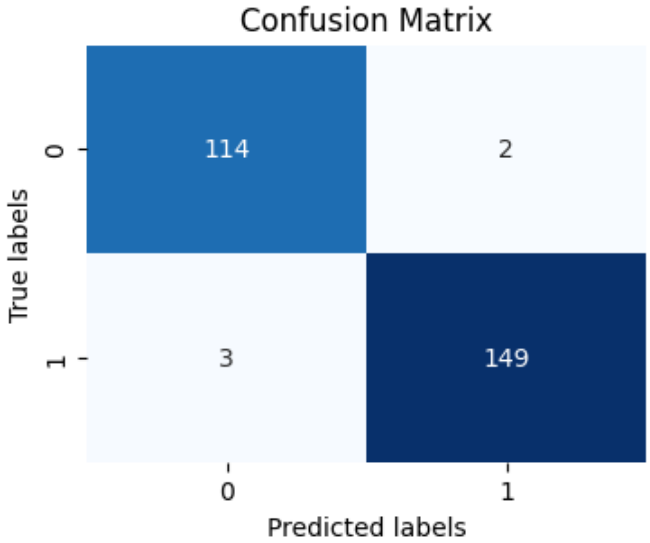
$$\hat{y_{\text{Visceral Fat}}} = 1.130678x - 19.646386$$

Berikut ini adalah plot setelah mengganti nilai mayoritas dengan pembulatan hasil prediksi.



2. Cholesterol Total (mg/dL) Karena data dari cholesterol total ini tidak merata, maka kita akan membagi data menjadi dua, yaitu data dengan kolesterol total sebesar 187 dan bukan 187.

F. Machine Learning Modelling. Setelah melakukan EDA, pertama-tama buatlah model yang mengklasifikasi apakah nilai Cholesterol Total sama dengan 187 atau bukan. Dengan *XGBClassifier* dilakukan CV dengan jumlah lipatan 5 didapatkan akurasi sebesar 0.9880485214377 dengan *confusion matrix* sebagai berikut.



Dengan demikian dapat disimpulkan bahwa model sudah dapat dengan baik memeriksa apakah kadar kolesterol total seseorang sama dengan 187 atau bukan.

Lanjut dengan model regresi, dicoba beberapa model dengan CV jumlah lipatan 5, didapatkan hasil sebagai berikut

RMSE dari CV Regresi

Model	RMSE
XGBRegressor	32.55845011260976
CatBoostRegressor	32.21672503030186
LGBMRegressor	32.53727556023499
RandomForestRegressor	31.52978966740082
LinearRegression	32.2936294706871
Ridge	32.28843252235978
Lasso	32.26098747866248
LogisticRegression	39.77495361746239

Dapat dilihat bahwa model *Random Forest* memiliki skor RMSE paling rendah dibandingkan dengan model-model lainnya.

Selanjutnya setelah memilih fitur yang tidak memiliki dampak signifikan terhadap kadar kolesterol, yaitu 'IMT (kg/m2)', 'Tinggi badan (cm)', 'Lingkar perut (cm)', 'Masa Kerja', 'Tekanan darah (S)', 'Tekanan darah (D)' berdasarkan *p-value*.

Untuk model yang dibuat menggunakan gabungan 2 jenis algoritma, yaitu *XGBClassifier* dan *RandomForestRegressor*. Pertama-tama akan menjalankan terlebih dahulu *classifier*, jika hasil prediksi dari *classifier* sama dengan 0, maka akan lanjut ke regresi. Pada regresi, baru dilakukan *drop* fitur yang tidak memiliki korelasi terhadap target. Setelah itu, gabungkan hasil keduanya, jika hasil klasifikasi sama dengan 1 maka akan diubah menjadi 187.

Untuk pengujian model yang sudah dibuat, akan dilakukan *Stratified K Fold* dengan jumlah lipatan 5 dan data sudah terdistribusi dengan proposi yang seimbang (train 0.8 dan test 0.2). Didapatkan skor RMSE dari masing-masing fold sebagai berikut [22.61224077432002, 21.426462227397447, 20.60937332032967, 15.738653814116283, 20.87801201665978] dengan rata-rata 20.252948430564636.

Supplementary Note 1: Penutup

A. Kesimpulan dan Saran. Model ini dibuat dari data yang diberikan, yaitu dengan data yang *skewed* karena mayoritas memiliki nilai total fat sebesar 187. Hal ini memaksa kami untuk memproses data dengan dua langkah, yakni classifier terlebih dahulu untuk menentukan apakah nilainya 187 atau bukan, kemudian melakukan regression untuk non-187. Saran untuk penelitian selanjutnya adalah mencoba untuk melakukan survey untuk memperoleh data baru yang memiliki persebaran dari variabel target yang lebih baik.