PREDIKSI KOLESTEROL TINGGI MENGGUNAKAN LOGISTIC REGRESSION
TIM "INVARIANT"

DI SUSUN OLEH:

1. Shawn Michael Dayanti Intong
2. Sobirin Nur Imam

PROGRAM STUDI FISIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS INDONESIA

2024

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 LATAR BELAKANG

Penyakit kardiovaskular, termasuk penyakit jantung koroner dan stroke, masih menjadi ancaman dunia (global threat) dan merupakan penyakit yang berperan utama sebagai penyebab kematian nomor satu di seluruh dunia. Menurut data dari Organisasi Kesehatan Dunia (WHO), lebih dari 17 juta orang meninggal karena penyakit jantung dan pembuluh darah. Di Indonesia sendiri, angka kematian akibat penyakit kardiovaskular mencapai 651.481 penduduk setiap tahun, dengan stroke menyebabkan 331.349 kematian, penyakit jantung koroner 245.343 kematian, penyakit jantung hipertensi 50.620 kematian, dan penyakit kardiovaskular lainnya (IHME, 2019).

Salah satu faktor risiko utama untuk penyakit kardiovaskular adalah tingginya kadar kolesterol dalam darah. Kolesterol merupakan jenis lemak yang ditemukan dalam darah dan diperlukan untuk fungsi normal tubuh, seperti pembentukan membran sel, produksi hormon, dan produksi vitamin D. Namun, kadar kolesterol yang tinggi dapat menyebabkan penumpukan plak di dalam arteri, yang pada gilirannya dapat menyebabkan penyempitan pembuluh darah dan risiko serangan jantung atau stroke. Dalam rangka menintkatkan indeks kualitas Global Disease Burden (GDB) dan terwujudnya Tujuan Pembangunan Berkelanjutan (SDGs) dari suatu negara, kita harus bisa memprediksi faktor apa saja yang membuat kolesterol dalam darah menjadi tinggi. Dengan cara ini, maka akan membantu seseorang pasien yang memiliki kadar kolesterol tinggi dalam darah dapat dengan mudah untuk diprediksi dan dilakukan tindak lanjut medis, sehingga risiko penyakit berbahaya seperti kardiovaskular dan kematian dapat dicegah.

Terdapat beberapa faktor yang dapat menjadi acuan dalam dunia medis untuk memprediksi kolesterol tinggi, diantaranya adalah tekanan darah, kadar lemak HDL dan LDL, serta Trigliserida. Pada praktiknya, terkadang kita harus melakukan tes secara manual untuk meghitung nilai kolesterol total, yang membuatnya bisa menjadi sangat lama dan antrian panjang. Oleh karena itu, dibutuhkan sebuah solusi untuk pemercepat permasalahan ini. Salah satunya adalah dengan memanfaatkan perkembangan teknologi, yaitu dengan membuat model algoritma machine learning yang dapat mempredikisi nilai kolesterol total dan faktor apa saja yang mempengaruhinya.

1.2 RUMUSAN MASALAH

Berdasarkan uraian di atas, rumusan masalah yang di dapat adalah:

1. Bagaimana algoritma atau metode serta model machine learning yang dapat menentukan faktor apa saja yang mempengaruhi kadar kolesterol total di dalam darah?

2. Bagaimana mempredikisi seseorang yang dapat berisiko menderita penyakit Kardiovaskular berdasarkan kadar kolesterol total dan data kesehatan?

1.3 BATASAN MASALAH

Batasan masalah yang digunakan adalah:

1. Dataset yang digunakan adalah Data Kesehatan Karyawan – Data 2 yang disediakan oleh pihak MCF ITB.
2. Model machine learning yang digunakan adalah .....

1.4 TUJUAN PENELITIAN

Tujuan penelitian ini adalah:

1. Membuat metode atau algoritma model machine learning yang dapat menentukan faktor apa saja yang mempengaruhi kadar kolesterol total di dalam darah.
2. Memprediksi seseorang yang dapat berisiko menderita penyakit kardiovaskular berdasarkan kadar kolesterol total dan data kesehatan

1.5 MANFAAT PENELITIAN

Manfaat penelitian ini adalah mengetahui faktor apa saja yang mempengaruhi kadar kolesterol total dan prediksi risiko seseorang dapat menderita penyakit kardiovaskular menggunakan model prediksi machine learning yang efektif dan efisiensi waktu.

1.6 METODOLOGI PENELITIAN

Metodologi penelitian menjelaskan mengenai jenis penelitian, jenis dan sumber data, metode analisis data, tahapan penelitian, proses pembuatan program dan flowchart program. Tahapan penelitian terdiri dari tahapan pendahuluan, tahapan studi pustaka, tahapan pengolahan data, interpretasi hasil dan tahapan kesimpulan dan saran. Berikut merupakan penjelasan dari tahapan penelitian :

1. Tahapan pendahuluan. Tahapan ini terdiri dari menentukan topik, identifikasi dan perumusan masalah, menentukan tujuan penelitian, dan menentukan batasan serta metodologi penelitian.
2. Tahapan studi pustaka. Tahapan ini terdiri dari melakukan studi pustaka dari literatur yang berkaitan dengan kardiovaskular, kolesterol total, kadar HDL dan LDL, visceral

fat, machine learning, data demografi penyakit di Indonesia, serta literatur terkait lainnya.

3. Tahapan pengumpulan dan pengolahan data. Tahapan ini terdiri dari pengumpulan data dari dataset yang diberikan pihak MCF ITB, analisis data/data understanding, perancangan Algoritma, dan melakukan pengujian model, serta menentukan waktu hasil training dan testing.
4. Tahapan interpretasi hasil. Tahapan ini terdiri dari tahapan interpretasi hasil berdasarkan, perancangan Algoritma machine learning, dan analisis hasil training dan testing.
5. Tahapan kesimpulan dan saran. Merupakan tahapan terakhir dari penelitian.

BAB II
PEMBAHASAN

2.1 ANALISIS

- **Perbedaan Kolesterol dan Trigliserida**

  Berdasarkan studi menurut Clinical Methods: The History Physical and Laboratory Examinations dalam National Center for Biotechnology Information, kolesterol adalah sejenis zat lilin yang memiliki fungsi penting dalam membangun sel dan memproduksi hormon estrogen dan progesteron, vitamin D dan asam empedu untuk pencernaan. Trigliserida merupakan zat yang secara eklusif berasal dari lemak dari makanan yang kita konsumsi. Kelebihan kalori dan gula juga akan diubah oleh tubuh menjadi trigliserida disimpan dalam bentuk lemak di seluruh tubuh kita seperti di perut, paha, di bawah kulit dan bagian lainnya.

Kolesterol merupakan lemak yang diproduksi oleh tubuh dari makanan yang telah dikonsumsi, ini bisa didapatkan dari makanan yang mengandung lemak jenuh. Perlu diketahui bahwa Kolesterol tidak dapat larut di dalam tubuh kita, sehingga akan mengalir dan berputar terus dan bergabung dengan protein membentuk lipoprotein. Lipoprotein ini membentuk kolesterol menjadi dua jenis yaitu HDL dan LDL. HDL berfungsi untuk membersihkan kolesterol yang ada di dalam tubuh kita, termasuk di dalam darah dan membawanya kembali ke bagian hati. LDL berperan membawa kolesterol dari hati ke berbagai organ, LDL menjadi jahat jika jumlahnya menjadi sangat banyak, karena akan membawa tumpukan lemak ke berbagai bagian tubuh termasuk ke dalam pembuluh darah yang menyebabkan penyumbatan.

- **Kolesterol Total (Cholesterol Total / CT)**

  Kolesterol total adalah ukuran dari total kolesterol dalam darah, yang mencakup kolesterol baik (HDL), kolesterol jahat (LDL), dan trigliserida. Kolesterol HDL sering disebut sebagai "kolesterol baik" karena memiliki efek protektif terhadap kesehatan jantung, sedangkan kolesterol LDL sering disebut sebagai "kolesterol jahat" karena penumpukannya dalam arteri dapat menyebabkan aterosklerosis. Ada beberapa faktor gaya hidup yang dapat menyebabkan tingginya kolesterol, antara lain merokok, mengonsumsi alkohol secara berlebihan, kurang bergerak, dan mengalami stres. Merokok dapat menurunkan kadar kolesterol baik (HDL) dan meningkatkan kadar kolesterol jahat (LDL), sementara alkohol berlebihan dapat meningkatkan total kolesterol dalam tubuh. Kurangnya aktivitas fisik juga dapat menyebabkan tubuh tidak menghasilkan cukup kolesterol baik. Selain itu, tekanan pikiran yang tinggi dapat memicu perubahan hormonal yang menyebabkan tubuh menghasilkan kolesterol. Selain faktor gaya hidup, faktor genetik juga berperan dalam meningkatkan risiko kolesterol tinggi. Konsumsi makanan tertentu juga dapat memengaruhi kadar kolesterol, sehingga terkadang perubahan pola makan diperlukan untuk mengatasi masalah kolesterol. Kolesterol Total dapat diukur menggunakan rumus perhitungan sebagai berikut:

  Kolesterol Total = Jumlah LDL + HDL + 1/5Trigliserida

- **Hubungan Kolesterol Total dengan Tempat Lahir**

Penyakit jantung sering disebabkan oleh faktor keturunan. Selain faktor keturunan, risiko penyakit jantung meningkat bila ada keluarga yang memiliki faktor risiko penyakit jantung, seperti: mengalami tekanan darah tinggi, diabetes dan kolesterol tinggi. Faktor genetik tersebut dapat memengaruhi risiko penyakit jantung, salah satunya mengatur sistem kardiovaskular. Variasi dapat menyebabkan tubuh memproses kolesterol secara berbeda, mengubah kekuatan otot pembuluh darah, dll.

Berdasarkan data Riskesdas 2018, ditemukan bahwa Prevalensi Penyakit Jantung berdasarkan diagnosa dokter di Indonesia mencapai 1,5%, dengan angka prevalensi tertinggi tercatat di Provinsi Kalimantan Utara sebesar 2,2%, DIY 2%, dan Gorontalo 2%. Di samping ketiga provinsi tersebut, terdapat 8 provinsi lainnya dengan prevalensi yang lebih tinggi daripada rata-rata nasional. Delapan provinsi tersebut mencakup Aceh (1,6%), Sumatera Barat (1,6%), DKI Jakarta (1,9%), Jawa Barat (1,6%), Jawa Tengah (1,6%), Kalimantan Timur (1,9%), Sulawesi Utara (1,8%), dan Sulawesi Tengah (1,9%). Selain itu, penduduk perkotaan juga lebih rentan terhadap Penyakit Jantung dengan prevalensi sebesar 1,6% dibandingkan dengan penduduk di pedesaan yang hanya 1,3%.

Berdasarkan acuan faktor keturunan genetik dan data Riskesdas 2018, kita menggunakan data tempat lahir dan data wilayah provinsi yang ada di Indonesia untuk menganalisanya. Tempat lahir juga bisa mempengaruhi lingkungan yang tidak sehat bagi penderita penyakit jantung yang disebabkan oleh kolesterol tinggi, misalnya seseorang dilahirkan di tempat yang memiliki kebiasaan buruk dengan sering memakan makanan berkolesterol tinggi, maka faktor ini bisa menjadi penyebab penyakit jantung.

- **Hubungan Kolesterol Total dengan Kadar Gula Darah Puasa**

Glukosa adalah gula yang terkandung di dalam darah karena adanya proses pencernaan karbohidrat dari makanan yang dikonsumsi. Secara umum, kadar glukosa atau gula darah ini menjadi salah satu indikator yang diperiksa dalam skrining rutin hingga membantu menegakkan diagnosis diabetes. Terdapat beberapa jenis pemeriksaan kadar gula darah yang digunakan untuk mendiagnosis diabetes, salah satunya adalah tes gula darah puasa. Sebagai informasi, gula darah puasa merupakan kadar glukosa di dalam darah setelah tidak mengonsumsi makanan atau minuman (kecuali air putih) dalam kurun waktu tertentu.

Diabetes mellitus (baik tipe 1 maupun tipe 2) meningkatkan risiko seseorang terkena penyakit arteri koroner dan arteri perifer. Diabetes berhubungan dengan penurunan kadar HDL dan peningkatan kadar trigliserida serta LDL. Sekitar 7 dari 10 orang dengan diabetes tipe 2 didiagnosis dengan dislipidemia yang terkait dengan diabetes, yang mencakup tingginya kadar trigliserida, kadar LDL "kecil padat" yang tinggi, dan kadar HDL yang rendah. LDL "kecil padat" adalah jenis protein kolesterol tertentu yang dapat dengan mudah masuk ke dalam dinding arteri dan menyebabkan kerusakan. Kehadiran terlalu banyak LDL "kecil padat" dalam darah dapat menyebabkan pertumbuhan plak. Para peneliti terus meneliti keterkaitan antara diabetes dan penyakit jantung. Dengan informasi tersebut kita bisa melakukan analisa hubungan antara kadar kolesterol total dengan kadar gula darah puasa, jika kadar gula darah puasa tinggi, hal ini dapat terindikasi juga memiliki kadar kolesterol tinggi.

- **Hubungan Kolesterol Total dengan Visceral Fat**

Faktor obesitas juga menjadi salah satu faktor penyebab kolesterol tinggi. Obesitas yang terjadi karena penumpukan lemak di tubuh menjadi salah satu ciri bahwa seseorang memiliki kolesterol tinggi (dr. Primarini, Halosehat). Dengan rincian tersebut, kita dapat menggunakan data Visceral fat dan lingkar perut dari seseorang untuk memprediksi apakah seseorang memiliki kolesterol tinggi. Visceral fat adalah lemak perut yang terdapat di sekeliling organ vital seperti jantung, hati, dan ginjal. Berbeda dengan lemak subkutan yang terletak di bawah kulit, visceral fat tidak terlihat dari luar. Dengan tingginya Visceral fat ini, maka lingkar perut akan menjadi lebar, karena terdapat penumpukan lemak di area perut. Kita dapat menguji hubunganya dengan melakukan regresi antara data lingkar perut atau visceral fat dengan data kolesterol total.

- **Hubungan Kolesterol Total dengan Indeks Massa Tubuh (IMT)**

Menurut Riset Kesehatan Kesehatan Dasar 2013, prevalensi obesitas pada penduduk laki-laki dewasa terus meningkat dari tahun sebelumnya. Pada tahun 2013 sebanyak 19,7 persen mengalami peningkatan dari tahun 2007 (13,9%) dan tahun 2010 (7,8%). Begitu pula prevalensi obesitas pada perempuan dewasa (>18 tahun), padatahun 2013 sebanyak 32,9 persen, naik 18,1 persen dari tahun 2007 (13,9%) dan 17,5persen dari tahun 2010 (15,5%). Penelitian MONICAI atau Multinational

Monitoring of Trends Determinants in Cardiovascular Diseases menyatakan bahwa penambahan berat badan dapat diiringi dengan peningkatan serum kolesterol. Setiap peningkatan 1 kg/m2 indeks massa tubuh (IMT) berhubungan dengan kolesterol totalplasma 7,7 mg/dl dan penurunanHDL 0,8 mg/dl.

- **Hubungan Kolesterol Total dengan Jenis Kelamin**

  Prevalensi Penyakit Jantung Koroner (PJK) cenderung lebih tinggi pada perempuan (1,6%) daripada laki-laki (1,3%). Penyakit jantung koroner adalah penyakit yang juga disebabkan oleh kadar kolesterol yang tinggi. Keadaan di mana kadar kolesterol sudah melebihi kadar normal di dalam darah disebut dengan Hiperkolesterolemia. Angka kejadian hiperkolesterolemiadi Indonesia untuk pria sebesar 11,4% dan untuk wanita sebesar 13,4%. Berdasarkan data tersebut, kita dapat menguji kadar kolesterol total dari setiap jenis kelamin apakah memiliki pengaruh yang berbeda, karena kadar kolesterol total juga memiliki korelasi dengan prevalensi penyakit Jantung Koroner.

- **Business Understanding**

  Berdasarkan dataset yang diberikan, dataset merupakan hasil survei kesehatan karyawan perusahaan. Kami berasumsi bahwa perusahaan sedang menerapkan langkah konkrit untuk pengembangan bisnis dan optimalisasi kinerja perusahaan, yaitu dengan melakukan survei mengenai kesehatan karyawan dengan tujuan perusahaan dapat menerapkan langkah preventif untuk kemajuan perusahaan. Jika karyawan dalam perusahaan tersebut tergolong sehat maka perusahaan dapat menggunakan dana perusahaan untuk kebutuhan pengembangan bisnis perusahaan. Akan tetapi jika hasil survei dan analisis data menunjukann bahwa karyawan banyak yang sakit atau berpotensi sakit maka perusahaan harus bisa memfasilitasi kesehatan karyawan demi menjaga kinerja perusahaan.

  Dalam melakukan langkah preventif tersebut, maka perusahaan harus bisa mengetahui data kesehatan dari karyawannya, kemudian mengetahui faktor apa saja yang mempengaruhi dari kesehatan tersebut, lalu mencari langkah yang solutif untuk menangani permasalahan tersebut. Untuk melakukan optimalisasi kesehatan karyawan tentu perusahaan tidak bisa mendukung pencegahan semua penyakit yang mungkin dapat dialami oleh karyawan. Namun, perusahaan dapat mencegah karyawan terkena

penyakit berbahaya yang paling sering terjadi dan menyebabkan kematian pada seseorang terhadap aktivitasnya di perusahaan.

Pekerjaan kantor yang memakan waktu seharian dan terkadang sampai lembur dapat meningkatkan risiko penyakit jantung dan hipertensi (tekanan darah tinggi). Sebuah penelitian menunjukkan bahwa orang yang bekerja lebih dari 10 jam sehari memiliki risiko 60% lebih besar terkena penyakit kardiovaskular. Ketika seseorang begadang, mereka sering kali mengonsumsi makanan tinggi gula atau kafein untuk tetap terjaga selama jam kerja. Kemudian, saat lapar di tempat kerja, banyak karyawan memilih makanan cepat saji dan berlemak agar cepat merasa kenyang. Selain itu, stress, kurang istirahat, dan jarang berolahraga juga menjadi pemicu kedua penyakit tersebut.

Menurut data WHO, penyakit kardiovaskular adalah penyakit yang paling banyak menyebabkan kematian dan membutuhkan penanganan serius serta terkadang gejalanya jarang terlihat. Penyakit kardiovaskular adalah penyakit yang disebabkan oleh kadar LDL dalam kolesterol total yang tinggi. Berdasarkan data kesehatan yang diperoleh dari survei karyawan, kita dapat mengolahnya untuk meminimalisir terjadinya penyakit kardiovaskular dengan cara mengetahui penyebabnya serta memprediksinya. Oleh karena itu, diperlukan sebuah model machine learning yang dapat mengetahui faktor apa saja yang mempengaruhi serta memprediksi tingginya kadar kolesterol total seseorang.

2.2 DATA CLEANING

Data cleaning merupakan proses pembersihan data kotor menjadi data bersih sehingga siap dipakai untuk melatih model. Data kotor sendiri dapat memiliki banyak arti seperti adanya pencilan atau outliers, data inkonsisten, missing values, dan lain-lain. Namun sebelum itu penting untuk mendapatkan gambaran atau summary dari data yang kita miliki itu sendiri. Dataset yang ada merupakan data sekunder dari panitia MCF DSC 2024 yang berisi data sampel hasil survey kesehatan suatu perusahaan sebanyak 1336 pegawai dan 15 variabel yang diukur. Pada Gambar 1, diperlihatkan lima data pertama dari dataset.

	Jenis Kelamin	Usia	Tekanan darah (S)	Tekanan darah (D)	Tinggi badan (cm)	Berat badan (kg)	IMT (kg/m2)	Lingkar perut (cm)	Glukosa Puasa (mg/dL)	Cholesterol Total (mg/dL)	Trigliserida (mg/dL)	Fat	Visceral Fat	Masa Kerja	Tempat lahir
0	M	19.0	126.0	88.0	172.5	49.50	16.53	66.0	84.0	187.0	99.0	26.4	6.0	0.97	Purworejo
1	M	19.0	120.0	80.0	158.0	53.60	21.50	71.0	84.0	187.0	99.0	26.4	6.0	0.60	Bogor
2	M	19.0	120.0	80.0	170.0	59.50	20.59	80.0	80.0	187.0	99.0	26.4	6.0	1.37	bandung
3	F	19.0	100.0	70.0	149.0	45.10	20.31	62.0	81.0	187.0	99.0	30.5	3.5	1.00	Jakarta
4	M	19.0	110.0	70.0	171.6	62.40	21.19	78.0	84.0	187.0	99.0	26.4	6.0	4.00	Teluk Betung

Gambar 1. Lima Data Pertama dari Dataset

Kemudian untuk tipe data variabel yang ada pada dataset sendiri terdiri atas variabel kategorik baik nominal ataupun ordinal dan juga variabel numerik. Perbedaan tipe data ini penting untuk diketahui mengingat cara preprocessing-nya yang berbeda. Pada Gambar 2, diperlihatkan tipe data masing-masing variabel.

```
Data columns (total 15 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Jenis Kelamin              1339 non-null   object
 1   Usia                       1339 non-null   float64
 2   Tekanan darah  (S)         1339 non-null   float64
 3   Tekanan darah  (D)         1339 non-null   float64
 4   Tinggi badan (cm)          1339 non-null   float64
 5   Berat badan (kg)           1339 non-null   float64
 6   IMT (kg/m2)                1339 non-null   float64
 7   Lingkar perut (cm)         1339 non-null   float64
 8   Glukosa Puasa (mg/dL)      1339 non-null   float64
 9   Cholesterol Total (mg/dL)  1339 non-null   float64
 10  Trigliserida (mg/dL)       1339 non-null   float64
 11  Fat                        1339 non-null   float64
 12  Visceral Fat               1339 non-null   float64
 13  Masa Kerja                 1339 non-null   float64
 14  Tempat lahir               1339 non-null   object
dtypes: float64(13), object(2)
memory usage: 157.0+ KB
```

Gambar 2. Tipe Data Variabel pada Dataset

Statistik deskriptif untuk masing-masing variabel juga penting untuk melihat bagaimana pemusatan serta persebaran data. Dengan itu dapat dicocokkan dengan fakta di lapangan. Hal ini yang diperlihatkan oleh Gambar 3.

	Usia	Tekanan darah (S)	Tekanan darah (D)	Tinggi badan (cm)	Berat badan (kg)	IMT (kg/m2)	Lingkar perut (cm)	Glukosa Puasa (mg/dL)	Cholesterol Total (mg/dL)	Trigliserida (mg/dL)	Fat	Visceral Fat	Masa Kerja
count	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000	1339.000000
mean	28.597461	113.147872	74.009709	164.940851	64.620500	23.693727	80.441972	84.571322	187.995519	106.982823	26.203510	6.231367	6.401837
std	4.767230	10.164592	7.718752	7.386617	12.799096	4.021585	10.688215	11.522057	21.104834	44.143456	3.678467	2.431923	4.554438
min	19.000000	80.000000	58.000000	138.500000	38.500000	14.850000	54.000000	65.000000	103.000000	34.000000	5.800000	0.500000	0.000000
25%	25.000000	110.000000	70.000000	160.000000	55.275000	20.855000	72.000000	84.000000	187.000000	99.000000	26.400000	6.000000	4.000000
50%	28.000000	110.000000	72.000000	165.000000	62.500000	23.200000	80.000000	84.000000	187.000000	99.000000	26.400000	6.000000	6.000000
75%	31.000000	120.000000	80.000000	170.000000	71.775000	26.000000	87.000000	84.000000	187.000000	99.000000	26.400000	6.000000	8.000000
max	39.000000	170.000000	100.000000	187.500000	139.750000	44.100000	128.000000	321.000000	308.000000	634.000000	40.900000	23.000000	31.000000

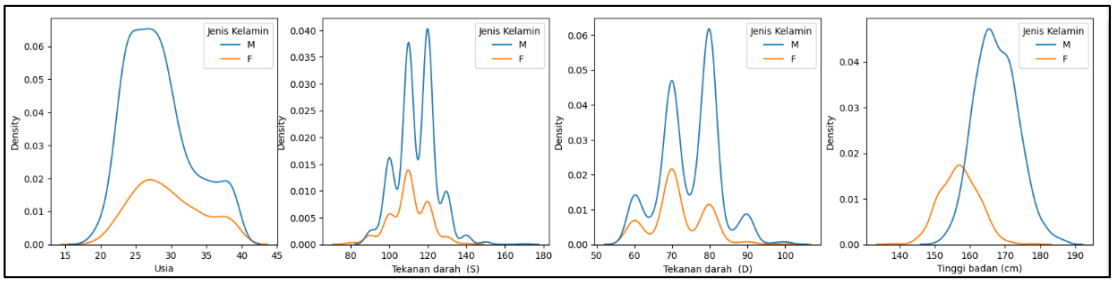
Gambar 3. Statistik Deskriptif Dataset

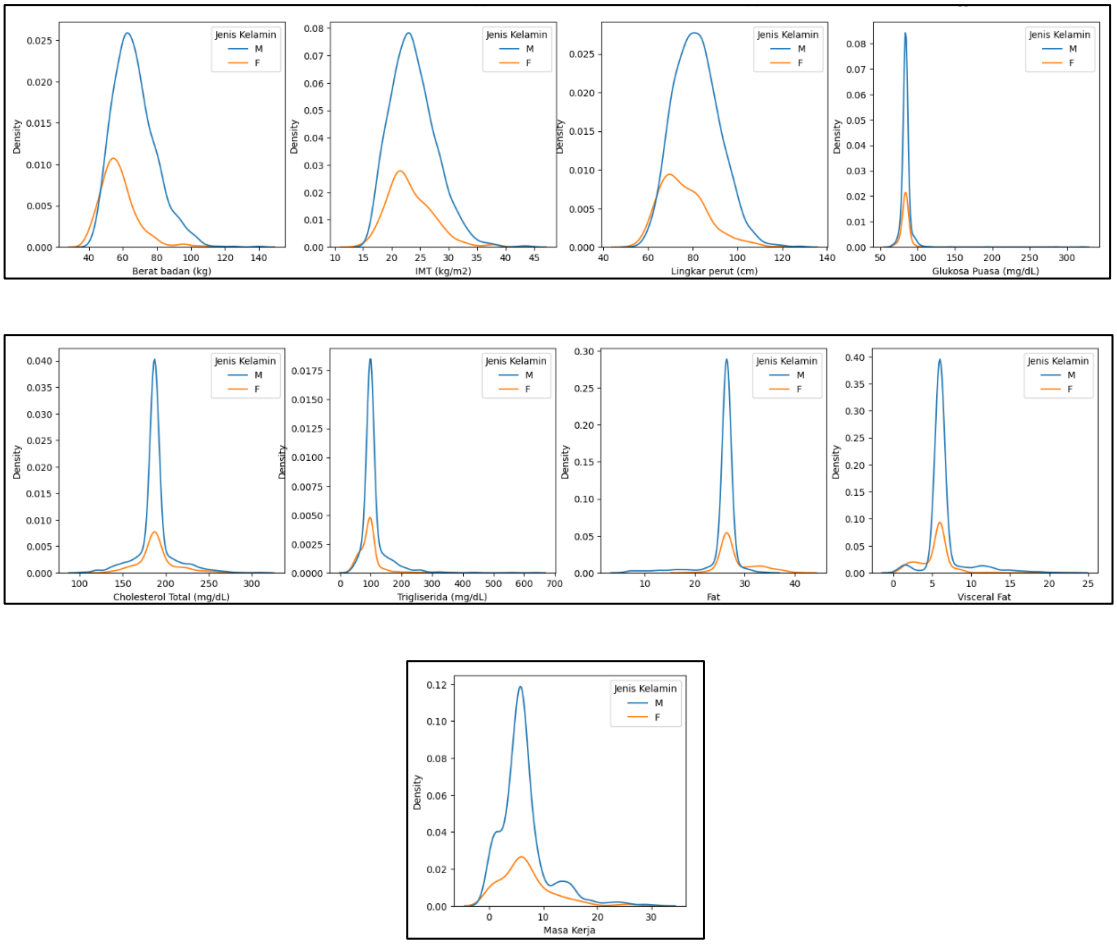
Berikutnya baru dapat kita lakukan pemeriksaan missing values. Setelah dilakukan pemeriksaan missing values, ternyata tidak ada nilai missing. Begitupun ketika dilakukan pemeriksaan data duplikat, tidak ada data yang duplikat. Hal ini berarti data yang diberikan di awal memang sudah cukup bersih.

```
Jenis Kelamin                0
Usia                         0
Tekanan darah  (S)           0
Tekanan darah  (D)           0
Tinggi badan (cm)            0
Berat badan (kg)             0
IMT (kg/m2)                  0
Lingkar perut (cm)           0
Glukosa Puasa (mg/dL)        0
Cholesterol Total (mg/dL)    0
Trigliserida (mg/dL)         0
Fat                          0
Visceral Fat                 0
Masa Kerja                   0
Tempat lahir                 0
dtype: int64
```

Gambar 4. Jumlah Missing Values Dataset

Visualisasi distribusi variabel numerik dataset kemudian dilakukan untuk mendapatkan gambaran bentuk distribusi data. Distribusi data sendiri penting untuk diketahui sebab dengan mengetahui distribusi data, terutama jika bentuknya menyerupai distribusi yang sudah umum dikenal contohnya adalah distribusi normal, kita dapat memahami dataset lebih baik dan mengetahui uji statistik yang tepat sehingga nantinya hasil model lebih akurat, reliable, dan valid. Hal ini yang diperlihatkan pada Gambar 5. Jika kita perhatikan sekilas, tidak ada satupun yang berbentuk distribusi normal walaupun ada beberapa variabel yang mendekati. Variabel seperti “Usia”, “Tekanan darah (S)”, “Tekanan darah (D)”, “Tinggi Badan (cm)”, “Berat Badan (kg)”, “IMT (kg/m2)”, dan “Lingkar Perut (cm)” memiliki distribusi right skewed dan platykurtik dilihat dari ekor distribusi yang tidak begitu tebal. Adapun variabel seperti “Glukosa Puasa (mg/dL)”, “Cholesterol Total (mg/dL)”, “Trigliserida (mg/dL)”, “Fat”, “Visceral Fat”, dan “Masa Kerja” memiliki distribusi simetrik dan leptokurtik dilihat dari ekor distribusi yang tebal.

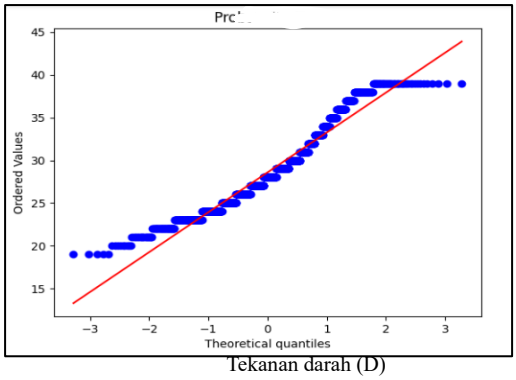




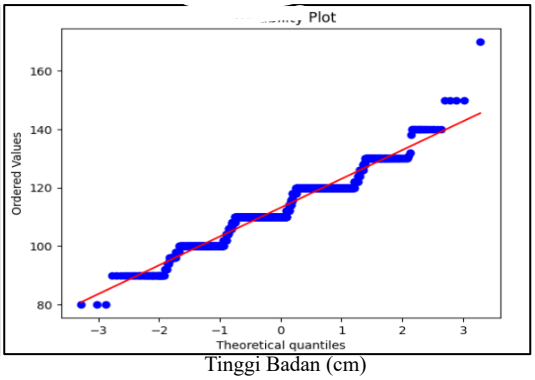
Gambar 5. Distribusi Vairabel Dataset

Adapun untuk memvalidasi pengamatan distribusi, dilakukan visualisasi Quantile-Quantile Plot (QQ-Plot) dimana semakin dekat titik dengan garis referensi maka variabel semakin mengikuti distribusi acuan atau asumsi. Hal ini dilakukan untuk menentukan apakah variabel pada dataset mengikuti suatu distribusi teoretis tertentu. Distribusi yang menjadi acuan atau pembanding disini adalah distribusi normal. Hasil ini diperlihatkan pada Gambar 6. Terlihat bahwa variabel yang benar-benar mengikuti distribusi normal adalah variabel "Tinggi badan (cm)". Variabel yang hampir berdistribusi normal namun right skewed seperti "Berat Badan (kg)", "IMT (kg/m2)", "Lingkar perut (cm)", dan "Masa Kerja". Adapun variabel "Usia", "Tekanan darah (S)", dan "Tekanan darah (D)" menunjukkan banyak puncak mengindikasikan distribusi berbentuk multimodal atau terdapat kelompok. Kemudian variabel seperti "Glukosa Puasa (mg/dL)", "Cholesterol Total (mg/dL)", "Trigliserida (mg/dL)", "Fat", dan" Visceral Fat" memiliki distribusi leptokurtik mengindikasikan terdapat banyak outlier.

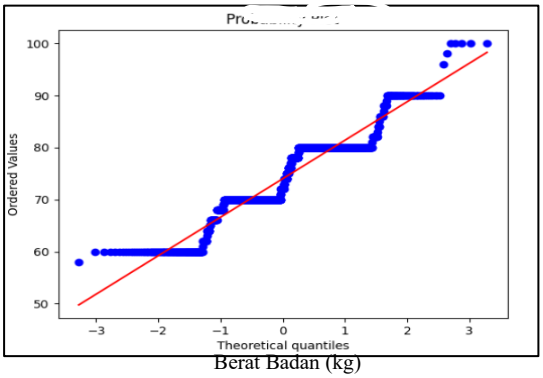
Usia Tekanan darah (S)



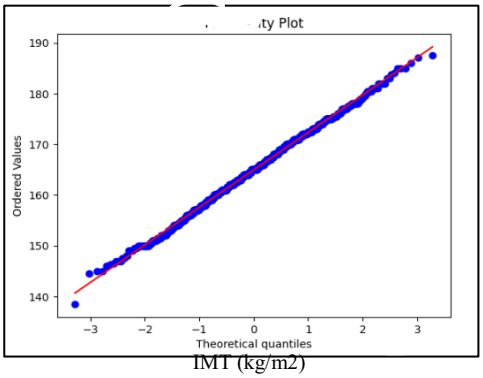
Tekanan darah (D)



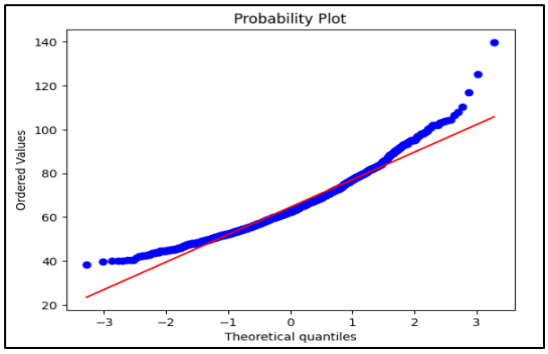
Tinggi Badan (cm)



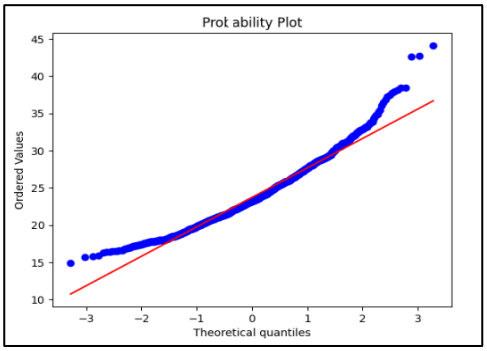
Berat Badan (kg)



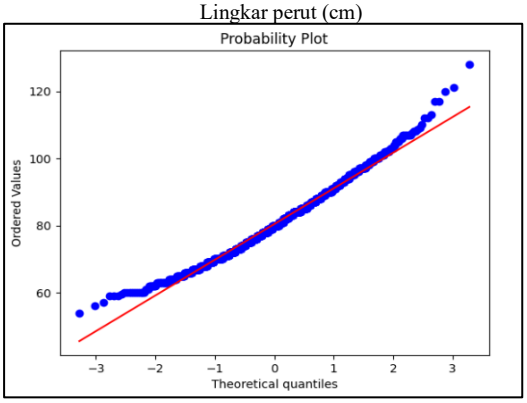
IMT (kg/m2)

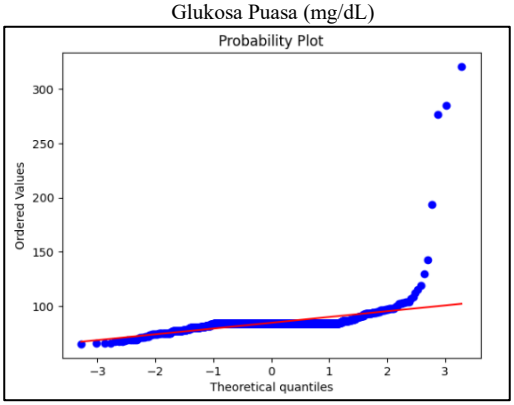


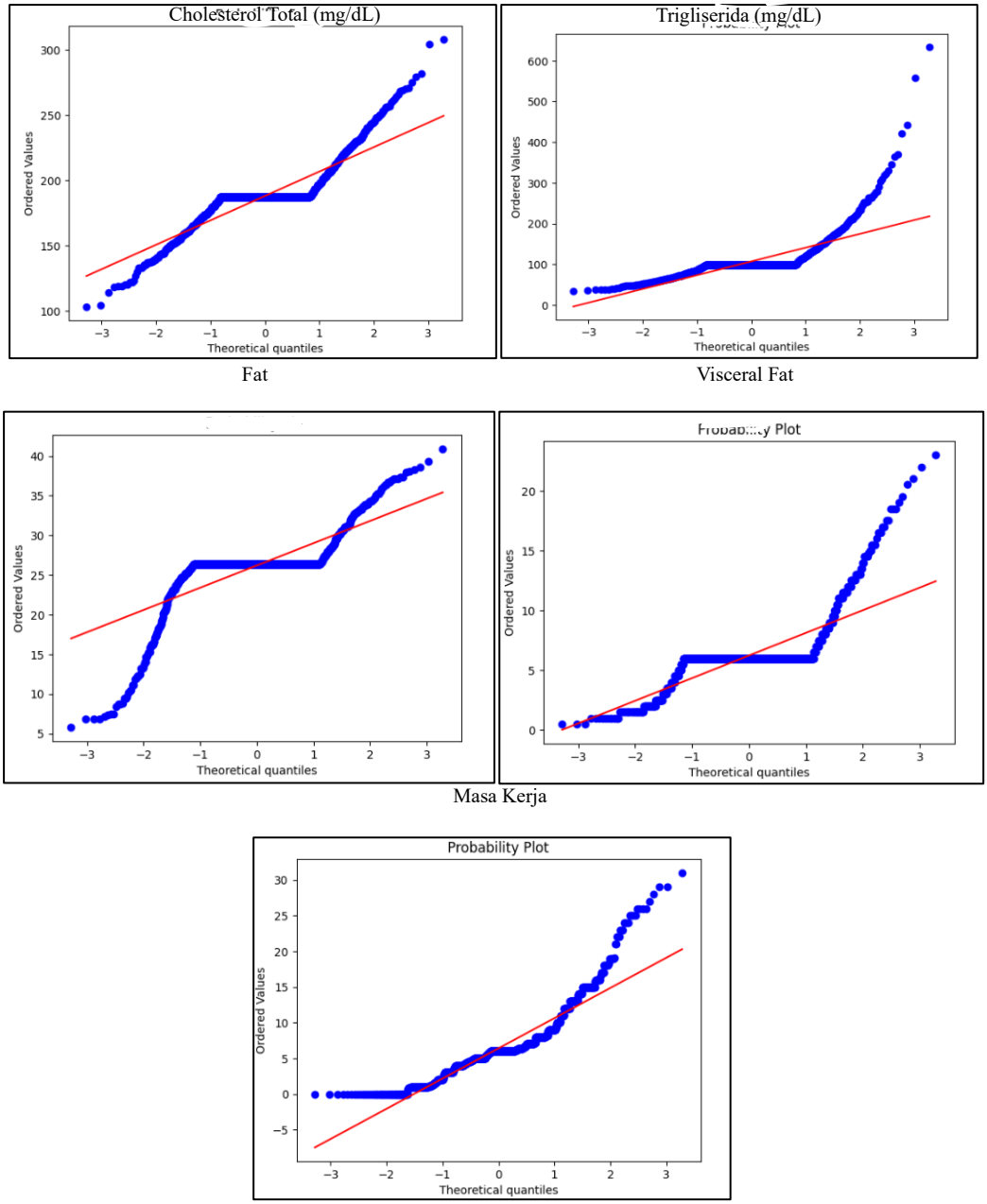
Lingkar perut (cm)



Glukosa Puasa (mg/dL)







Gambar 6. QQ Plot setiap Variabel

Terkahir kita melakukan preprocessing dataset. Tujuan preprocessing ini sendiri supaya dataset dapat digunakan untuk melatih model. Perlu diingat bahwa permasalahan machine learning-nya adalah klasifikasi. Model yang nantinya sudah dilatih akan digunkan untuk memprediksi apakah karyawan perusahaan tersebut berpotensi berkolesterol tinggi. Disini kita tetapkan bahwa karyawan dengan kolesterol total lebih dari atau sama dengan 200 mg/dL adalah berkolesterol tinggi sedangkan dibawah batas tersebut adalah normal. Selain itu beberapa variabel numerik kita ubah menjadi kategorik baik nominal maupun ordinal dengan tiap kategori memiliki rentang nilai tertentu. Hal ini dilakukan karena beberapa alasan antara lain noise dan outlier

pada variabel tersebut yang tinggi ataupun hasil iterasi kita yang sampai pada keputusan tersebut. Proses tersebut dinamakan binning atau diskritisasi. Referensi pembagian kelompok atau kategori variabel serta rentangnya tersebut kita sesuaikan dengan referensi terpercaya yang kita temukan baik dari jurnal maupun website kesehatan terkenal. Detail pembagiannya kita lampirkan pada lampiran.

Setelah melakukan binning, perlu diingat tipe datanya masih kategorik nominal maupun ordinal. Untuk mengubahnya menjadi numerik kita lakukan encoding. Variabel kategorik nominal dan ordinal memiliki perlakuan encoding yang berbeda. Perlu diingat bahwa variabel kategorik nominal tidak memiliki hirarki atau urutan antar kategorinya sehingga dapat kita lakukan dummy encoding atau one-hot encoding dengan menciptakan variabel baru sebanyak N-1 dengan N adalah jumlah kategori pada variabel parent tersebut. Variabel parent sendiri didefinisikan sebagai variabel sebelum dilakukan encoding. Variabel baru yang dibuat berlaku sebagai indikator kategori dari variabel parent-nya. Setelah itu variabel parent dapat di-drop atau dihilangkan dari dataset. Variabel "Jenis Kelamin" dan "Cholesterol Total (mg/dL)" kita encode menggunakan cara ini.

Adapun variabel kategorik ordinal yang memiliki hubungan hirarki atau urutan antar kategorinya. Variabel seperti ini menggunakan jenis ordinal encoding yang melakukan mapping kategori ke suatu nilai numerik yang unik dimana kategori yang memiliki posisi rendah di hirarki bersesuaian dengan nilai yang rendah pula begitupun sebaliknya. Hal ini dapat membuat model mempelajari hirarki yang ada pada variabel tersebut. Variabel "Usia", "IMT (kg/m2)", dan "Masa Kerja" kita encode menggunakan cara ini. Hasil encoding dapat dilihat pada Gambar 5 berikut.

	Jenis Kelamin	Usia	Tekanan darah (S)	Tekanan darah (D)	Tinggi badan (cm)	Berat badan (kg)	Lingkar perut (cm)	Glukosa Puasa (mg/dL)	Trigliserida (mg/dL)	Fat	Visceral Fat	Tempat lahir	CT	IMT	Senioritas
0	1	0	126.0	88.0	172.5	49.50	66.0	84.0	99.0	26.4	6.0	Purworejo	0	0	0
1	1	0	120.0	80.0	158.0	53.60	71.0	84.0	99.0	26.4	6.0	Bogor	0	1	0
2	1	0	120.0	80.0	170.0	59.50	80.0	80.0	99.0	26.4	6.0	bandung	0	1	0
3	0	0	100.0	70.0	149.0	45.10	62.0	81.0	99.0	30.5	3.5	Jakarta	0	1	0
4	1	0	110.0	70.0	171.6	62.40	78.0	84.0	99.0	26.4	6.0	Teluk Betung	0	1	1

Gambar 8. Hasil Encoding Dataset

Jika kita perhatikan dataset kembali, variabel "Tempat lahir" juga merupakan variabel kategorik nominal. Namun tidak kita encode layaknya variabel kategorik nominal yang lain seperti "Jenis Kelamin" dan "Cholesterol Total (mg/dL)". Hal ini

karena variabel tersebut terdapat sekitar 170 kategori. Artinya dengan melakukan dummy encoding atau one-hot encoding, kita akan membuat 169 variabel yang baru yang mayoritas berisi nilai nol. Tentu hal ini tidak diinginkan sebab selain akan memperlama trainning dan waktu komputasi, interpretasi pada model akhirnya akan sulit. Oleh karena itu, kita mencoba untuk mengelompokkan tempat lahir berdasarkan pulau terbesar terdekat dari tempat lahir tersebut. Pulau terbesar-nya sendiri yaitu pulau "Jawa", "Sumatera", "Kalimantan", "Sulawesi", dan "Papua". Dengan itu jumlah kategori akan tereduksi menjadi hanya 4 variabel dari yang tadinya 170 variabel. Namun karena pada dataset tidak terlalu banyak titik data dengan "Tempat lahir" di pulau Papua, maka cukup dibuat 3 variabel saja yaitu "Sumatera", "Kalimantan", dan "Sulawesi". Titik data sisa yang bersesuaian dengan "Tempat lahir" di Papua akan dikelompokkan ke pulau Sulawesi saja. Hasil encoding sesuai dengan Gambar 9.

	Jenis Kelamin	Usia	Tekanan darah (S)	Tekanan darah (D)	Tinggi badan (cm)	Berat badan (kg)	Lingkar perut (cm)	Glukosa Puasa (mg/dL)	Trigliserida (mg/dL)	Fat	Visceral Fat	CT	IMT	Senioritas	Sumatera	Kalimantan	Sulawesi
0	1	0	126.0	88.0	172.5	49.50	66.0	84.0	99.0	26.4	6.0	0	0	0	0	0	0
1	1	0	120.0	80.0	158.0	53.60	71.0	84.0	99.0	26.4	6.0	0	1	0	0	0	0
2	1	0	120.0	80.0	170.0	59.50	80.0	80.0	99.0	26.4	6.0	0	1	0	0	0	0
3	0	0	100.0	70.0	149.0	45.10	62.0	81.0	99.0	30.5	3.5	0	1	0	0	0	0
4	1	0	110.0	70.0	171.6	62.40	78.0	84.0	99.0	26.4	6.0	0	1	1	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1334	1	1	110.0	70.0	162.5	67.60	87.0	84.0	72.0	23.8	11.0	0	2	2	0	0	0
1335	0	1	120.0	70.0	150.0	60.50	77.0	84.0	105.0	38.1	9.0	0	2	2	0	0	0
1336	0	1	120.0	80.0	151.0	59.75	78.0	84.0	78.0	35.3	8.0	0	2	3	0	0	0
1337	1	1	110.0	70.0	166.2	57.00	69.0	84.0	98.0	17.1	5.0	1	1	1	0	0	0
1338	0	1	110.0	70.0	155.0	52.15	70.0	84.0	51.0	26.4	6.0	0	1	2	0	0	0

Gambar 9. Hasil Encoding "Tempat lahir"

Setelah melakukan preprocessing dataset, kita dapat membuat variabel-variabel baru yang diturunkan dari variabel-variabel yang sudah ada pada dataset. Proses ini dinamakan dengan rekayasa fitur/variabel atau feature engineering. Terdapat kemungkinan variabel baru yang dibuat dapat mempermudah pembelajaran model ketika proses trainning. Dua variabel pertama yang kita buat yaitu "Lintang" dan "Bujur" yang diturunkan dari "Tempat lahir" menggunakan bantuan library Geopy yang berfungsi mencari koordinat lokasi dengan alamat spesifik tertentu. Kemudian untuk variabel ketiga dan keempat, kita dapat melakukan transformasi koordinat dari kartesian ke polar dengan menggunakan set transformasi berikut,

$$r = \sqrt{x^2 + y^2}$$

$$\theta = \tan^{-1}\left(\frac{y}{x}\right)$$

dimana x dan y masing-masing adalah “Bujur” dan “Lintang” sehingga menghasilkan variabel baru yaitu “r” dan “θ” yang masing-masing merupakan jarak ke pusat koordinat dan posisi sudut yang diukur berlawanan jarum jam. Terakhir, kita dapat membuat variabel yang menangkap transformasi rotasi koordinat kartesian asalnya. Besar sudut rotasi yang kita gunakan sendiri adalah 15, 30, dan 45 derajat. Hasil feature engineering dapat dilihat pada gambar 10.

	Lintang	Bujur	r	theta	rot_15_x	rot_15_y	rot_30_x	rot_30_y	rot_45_x	rot_45_y
0	-7.707302	109.966512	110.236274	-0.069973	104.224697	21.016745	91.380142	48.308536	73.653403	72.213897
1	-7.707302	109.966512	110.236274	-0.069973	104.224697	21.016745	91.380142	48.308536	73.653403	72.213897
2	-7.707302	109.966512	110.236274	-0.069973	104.224697	21.016745	91.380142	48.308536	73.653403	72.213897
3	-7.707302	109.966512	110.236274	-0.069973	104.224697	21.016745	91.380142	48.308536	73.653403	72.213897
4	-7.707302	109.966512	110.236274	-0.069973	104.224697	21.016745	91.380142	48.308536	73.653403	72.213897
...	...	...	...	...	...	...	...	...	...	...
1328	3.484872	99.137631	99.198862	0.035137	96.661549	29.024834	87.598143	52.586803	73.777820	72.607698
1329	3.001910	99.083314	99.128778	0.030288	96.484084	28.544271	87.309622	52.141387	73.397242	72.221877
1330	-2.585530	115.384523	115.413487	-0.022404	110.783706	27.366282	98.633163	55.453127	81.172434	79.729304
1331	-8.033597	113.266224	113.550765	-0.070808	107.327523	21.555597	94.074629	49.675813	75.796292	74.312429
1332	-7.388571	112.291649	112.534463	-0.065703	106.553101	21.926406	93.553135	49.747134	75.551344	74.087293

Gambar 10. Hasil Feature Engineering

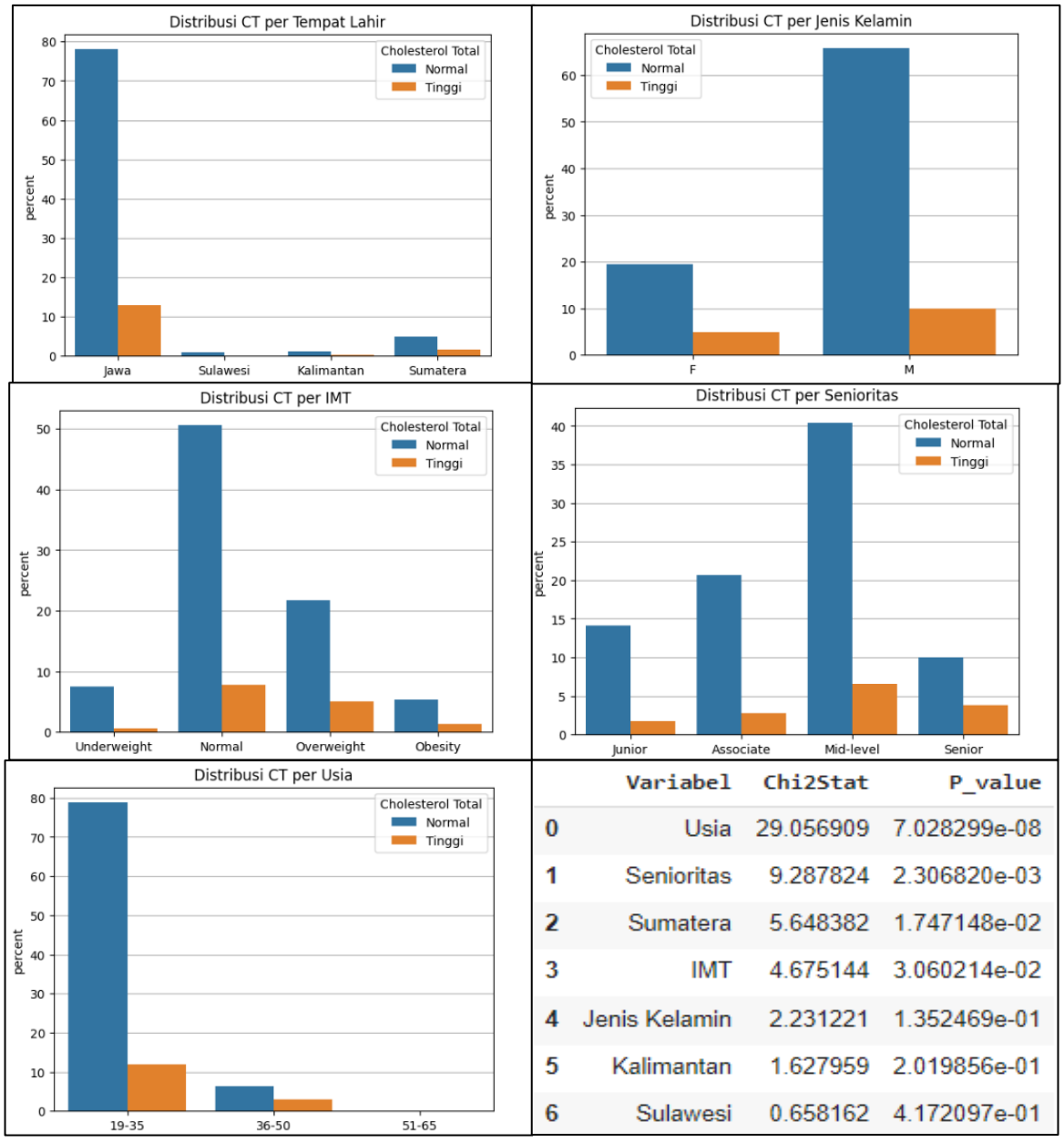
2.3 EXPLORATORY DATA ANALYSIS

Pada bagian ini, yaitu exploratory data analysis, dataset yang sudah diolah sedemikian rupa kita coba visualisasikan dan mencari pola dan hubungan antar variabel. Hal ini penting dalam proses seleksi variabel yang akan digunakan model supaya tidak ada multikolinearitas ataupun variabel yang redundant sehingga memperlama waktu trainning. Perlu diingat kembali bahwa variabel target pada permasalahan ini adalah “CT” yang memiliki tipe data kategorik nominal dengan kelas “0” yang artinya level kolesterol normal dan “1” yang artinya level kolesterol tinggi. Adapun variabel independen atau features tipe data kategorik yaitu “Jenis Kelamin”, “Usia”, “IMT”, dan “Senioritas”. Lalu variabel independen tipe data numerik yaitu “Tekanan darah (S)”, “Tekanan darah (D)”, “Tinggi badan (cm)”, “Berat badan (kg)”, “Lingkar perut (cm)”, “Glukosa Puasa (mg/dL)”, “Trigliserida (mg/dL)”, “Fat”, dan “Visceral Fat”.

Karena terdapat tipe data yang beragam jenis plot visualisasinya juga beragam. Untuk variabel kategorik dan target kategorik “CT”, plot yang digunakan adalah barplot yang membandingkan persentase antar kelas “CT” dan juga persentase distribusi variabel kategorik di setiap kelasnya. Adapun uji statistik chi squared yang dilakukan untuk menyelidiki apakah ada pengaruh atau hubungan signifikan antara variabel kategorik dan target “CT”. Asumsi uji statistik chi squared yaitu variabel yang dibandingkan adalah katergorik dan data independen. Hasil uji diperlihatkan pada Gambar 11.

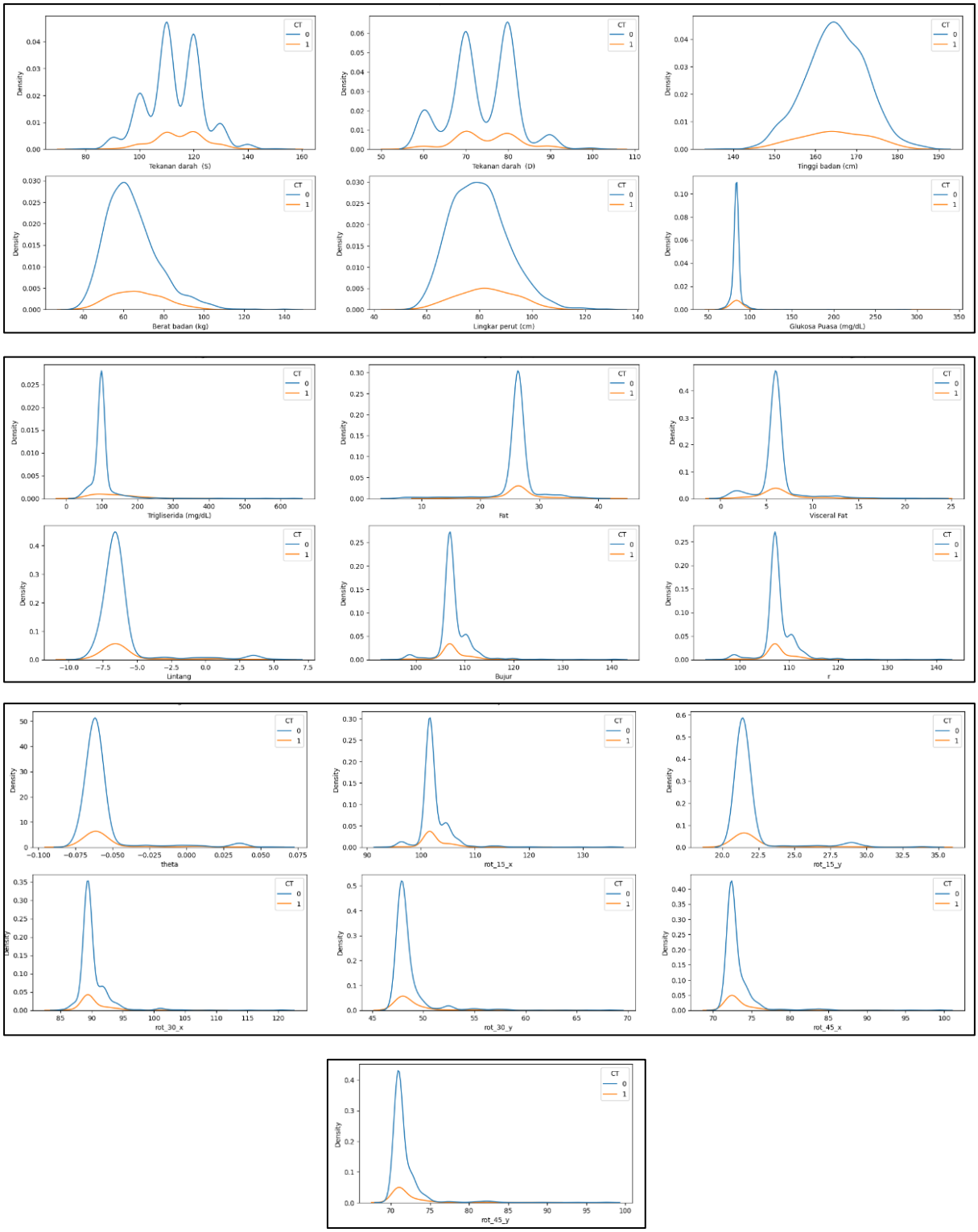
Kemudian untuk variabel numerik dan target kategorik “CT”, plot yang digunakan adalah kernel density plot. Dari jenis plot seperti ini, kita dapat melihat bagaimana perbedaan distribusi variabel numerik antar kelas “CT”. Jika sangat berbeda, maka terdapat kemungkinan bahwa variabel tersebut bukan merupakan prediktor yang baik bagi target “CT”. Namun inspeksi visual terkadang tidak cukup sehingga kita perlu melakukan uji statistik perbedaan distribusi antar kelas. Perlu diperhatikan bahwa pemilihan uji statistik bergantung pada distribusi masing-masing variabel. Sebelumnya sudah diketahui distribusi masing-masing variabel numerik dari QQ plot dan mayoritas tidak berdistribusi

normal. Oleh karena itu, uji statistik yang akan digunakan adalah uji statistik yang non-parametrik yaitu uji Kolmogorov-Smirnov. Hasil diperlihatkan pada Gambar 13.



	Variabel	Chi2Stat	P_value
0	Usia	29.056909	7.028299e-08
1	Senioritas	9.287824	2.306820e-03
2	Sumatera	5.648382	1.747148e-02
3	IMT	4.675144	3.060214e-02
4	Jenis Kelamin	2.231221	1.352469e-01
5	Kalimantan	1.627959	2.019856e-01
6	Sulawesi	0.658162	4.172097e-01

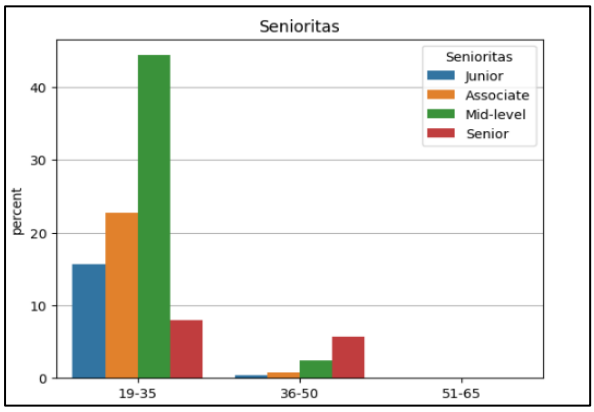
Gambar 11. Barplot dan Uji Chi Squared Variabel Kategork

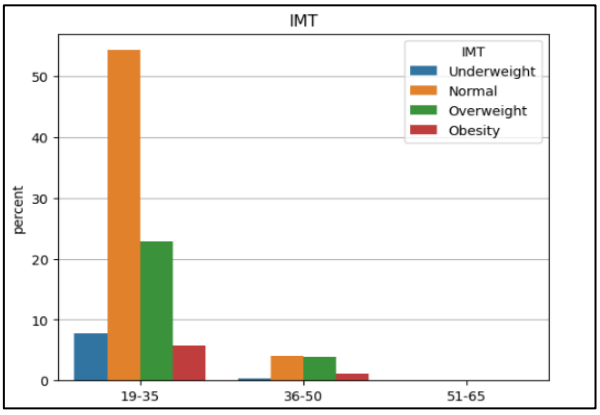


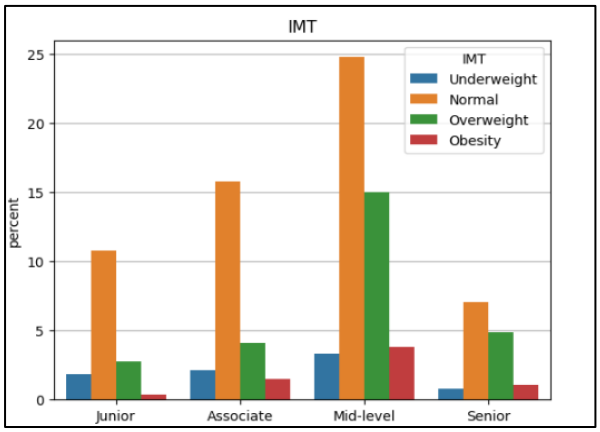
Gambar 12. Perbedaan Distribusi Variabel Numerik antar Kelas "CT"

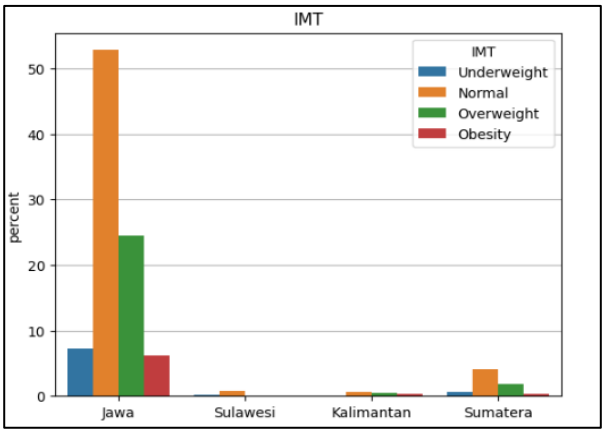
	Variabel	KSStat	P_value
0	Trigliserida (mg/dL)	0.515187	1.032772e-41
1	rot_15_y	0.192858	5.731980e-06
2	Berat badan (kg)	0.154439	5.572945e-04
3	Lintang	0.137663	2.979278e-03
4	theta	0.137018	3.165382e-03
5	Lingkar perut (cm)	0.130913	5.537441e-03
6	Fat	0.130530	5.728050e-03
7	Glukosa Puasa (mg/dL)	0.130121	5.940924e-03
8	rot_30_y	0.126107	8.448712e-03
9	Visceral Fat	0.125951	8.558203e-03
10	Tekanan darah (D)	0.112789	2.505768e-02
11	Tekanan darah (S)	0.104165	4.750047e-02
12	rot_45_x	0.083723	1.764841e-01
13	rot_45_y	0.083723	1.764841e-01
14	rot_30_x	0.071214	3.412243e-01
15	rot_15_x	0.069452	3.709426e-01
16	Tinggi badan (cm)	0.067517	4.054111e-01
17	r	0.055355	6.561059e-01
18	Bujur	0.055119	6.612795e-01

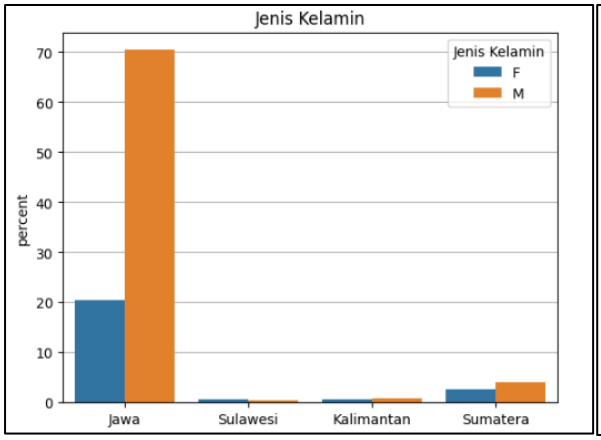
Gambar 13. Hasil Uji Kolmogorov-Smirnov

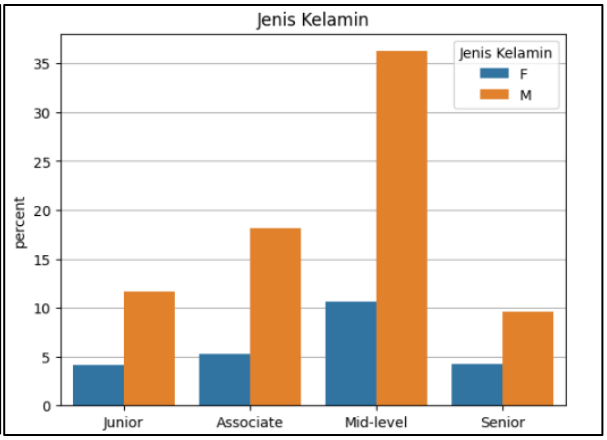


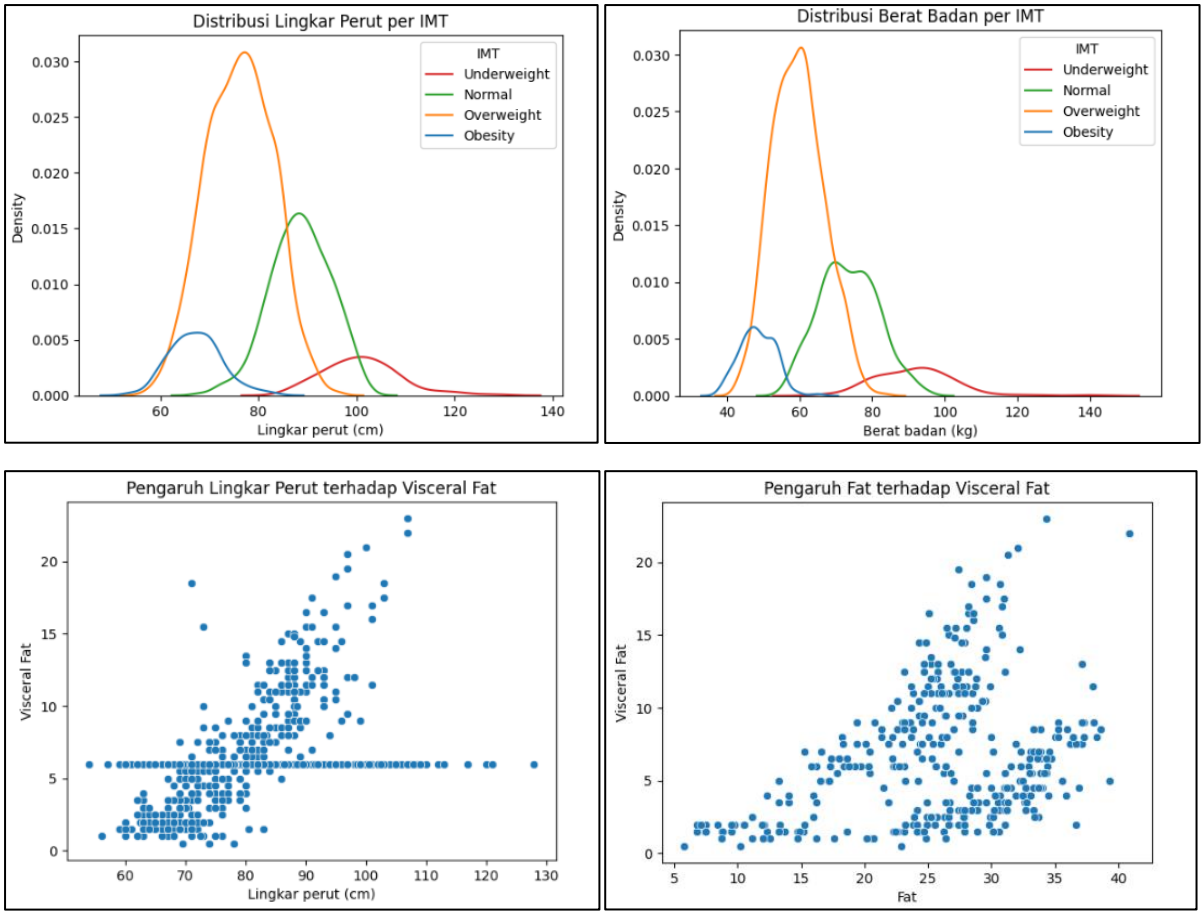




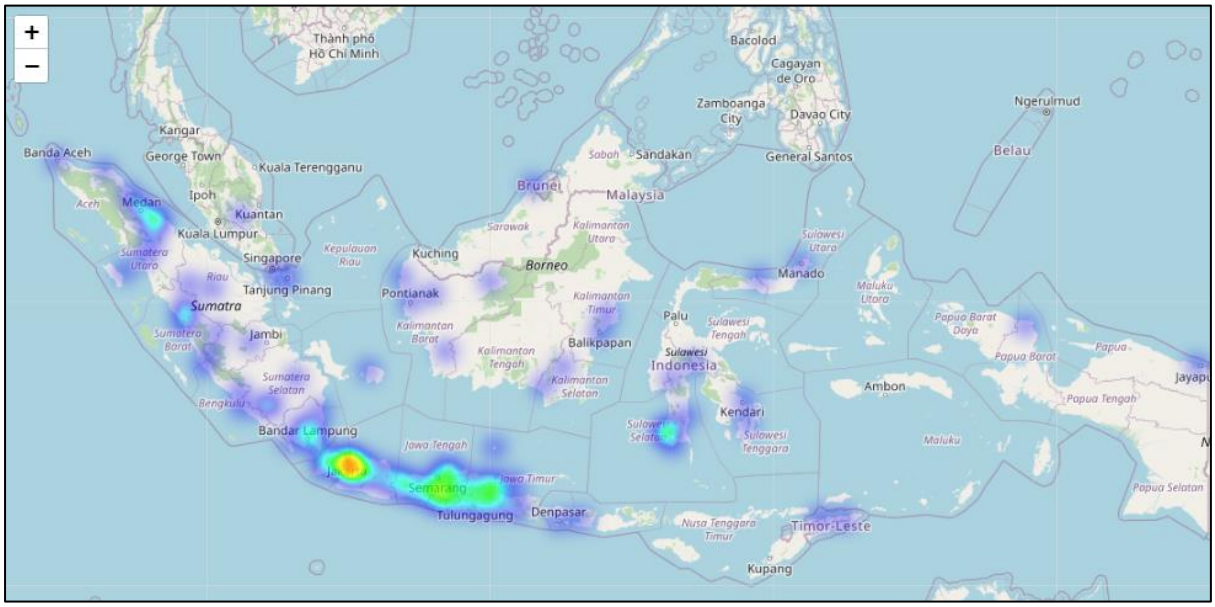








Gambar 14. Beberapa Visualisasi Tambahan



Gambar 15. Distribusi Tempat Lahir pada Peta

Berikut adalah beberapa hal yang dapat disimpulkan dari visualisasi dan uji statistik yang telah dilakukan sebelumnya:

a. Terdapat lebih banyak responden survey berjenis kelamin laki-laki dibandingkan perempuan. Walaupun begitu, terlihat bahwa dari uji chi squared tidak terdapat pengaruh atau hubungan signifikan antara “Jenis Kelamin” dan “CT” karena p value yang lebih besar dari nilai kritis yang kita asumsikan sebagai 0,05.
b. Responden yang mengisi survey mayoritas dengan level senioritas adalah “Mid-level” yang artinya karyawan tersebut sudah bekerja di perusahaan tersebut kurang lebih 5-10 tahun yang disusul oleh “Associate”, “Junior”, dan “Senior”. Terdapat pengaruh atau hubungan signifikan antara “Senioritas” dan “CT” karena p value yang lebih kecil dari nilai kritis yang kita asumsikan sebagai 0,05.
c. Responden yang mengisi survey mayoritas memiliki indeks massa tubuh yang “Normal” disusul oleh “Overweight”, “Underweight”, dan “Obesity”. Terdapat pengaruh atau hubungan signifikan antara “IMT” dan “CT” karena p value yang lebih kecil dari nilai kritis yang kita asumsikan sebesar 0,05.
d. Responden yang mengisi survey mayoritas berumur antara 19-35 tahun. Terdapat pengaruh atau hubungan signifikan antara “Usia” dan “CT” karena p value yang lebih kecil dari nilai kritis yang kita asumsikan sebesar 0,05.
e. Responden yang mengisi survey mayoritas memiliki tempat lahir atau berasal dari “Jawa” disusul oleh “Sumatera”, “Kalimantan”, dan “Sulawesi”. Terdapat hubungan atau pengaruh signifikan antara “Sumatera” dan “CT” karena p value yang lebih kecil dari nilai kritis yang kita asumsikan sebesar 0,05.
f. Terdapat perbedaan distribusi level “Senioritas” antar kelompok “Usia” dimana kelompok usia 19-35 didominasi oleh Mid-level kebawah sedangkan pada kelompok usia 36-50 didominasi oleh Senior.
g. Terdapat perbedaan distribusi “IMT” antar kelompok “Usia” dimana kelompok usia 19-35 didominasi oleh Normal sedangkan pada kelompok usia 36-50 didominasi oleh baik Normal maupun Overweight dan hampir tidak ada atau lebih jarang yang Underweight.
h. Tidak terdapat perbedaaan distribusi “IMT” antar baik level “Senioritas” maupun “Tempat lahir” dari responden walaupun terlihat bahwa responden dari pulau Sulawesi dan Kalimantan yang kurang.

i. Tidak terdapat perbedaan distribusi “Jenis Kelamin” baik berdasarkan “Tempat lahir” maupun level “Senioritas”.
j. Terdapat perbedaan distribusi variabel numerik baik untuk kelas “CT” bernilai 0 maupun 1 kecuali untuk variabel “rot_45_x”, “rot_45_y”, “rot_30_x”, “rot_15_y”, “Tinggi Badan (cm)”, “r”, dan “Bujur” karena nilai p value yang lebih besar dari nilai kritis yaitu 0,05.
k. Terdapat perbedaan distribusi “Lingkar Perut (cm)” dan “Berat Badan (kg)” terhadap “IMT” dimana responden yang Underweight cenderung memiliki lingkar perut dan berat badan yang lebih kecil. Sebaliknya, responden yang Obesity memiliki lingkar perut dan berat badan yang lebih besar.
l. Terdapat hubungan linear antara “Lingkar Perut (cm)” dan “Visceral Fat” yang artinya jika lingkar perut responden yang terukur bertambah besar maka visceral fat juga bertambah besar.
m. Terdapat hubungan linear antara “Fat” dan “Visceral Fat” dimana terlihat juga bahwa terdapat percabangan dengan kemiringan garis yang berbeda ketika nilai “Fat” sebesar 15 persen.
n. Distribusi tempat lahir responden berdasarkan peta berpusat di pulau “Jawa” terutama di daerah Jakarta, Bogor, Depok, Tanggerang, dan Bekasi (Jabodetabek). Walaupun demikian, masih terdapat responden di luar pulau “Jawa” yang jika diperhatikan masih cukup merata sehingga kita dapat menyimpulkan bahwa survey cukup representatif dan mewakili hampir seluruh kota/desa di Indonesia.

2.4 MACHINE LEARNING DEPLOYMENT

Setelah kita mengenal masing-masing variabel seperti distribusi beserta hubungannya dengan variabel lain, maka dataset dapat dipakai untuk membuat model. Tahap pembuatan model sampai deployment disebut juga dengan machine learning deployment. Pertama, kita perlu dengan jelas mendefinisikan variabel independen/X dan variabel dependen/y. Sesuai dengan uji Chi Squared untuk variabel kategorik pada bagian Exploratory Data Analysis, maka kita mengambil variabel independen/X dengan nilai $p_{value} < \alpha$ dimana nilai α kita ambil sebesar 0,05 yang merupakan nilai kritis uji sehingga X dan y diperlihatkan seperti pada Tabel 1.

Jenis Variabel	Nama Variabel
X	Trigliserida (mg/dL), Berat badan (kg), Tinggi Badan (cm), Glukosa Puasa (mg/dL), Lingkar perut (cm), Fat, Visceral Fat, Tekanan darah (D), Tekanan darah (S), Usia, Senioritas, IMT, Sumatera, rot_15_x, rot_15_y, rot_30_x, rot_30_y, rot_45_x, rot_45_y, Lintang, Bujur, r, dan theta
y	CT

Tabel 1. Variabel X dan y

Berikutnya setelah seleksi/pemilihan variabel independen/X, perlu dilakukan perhitungan Variance Inflation Factor (VIF) masing-masing variabel. Hal ini untuk mencegah terjadinya multikolinearitas yang dapat menyebabkan estimasi parameter pada model yang tidak konsisten. Untuk threshold yang akan digunakan sendiri adalah 15, yang artinya variabel dengan $VIF \geq 15$ memiliki multikolinearitas yang kuat sehingga akan dihilangkan dari model. Hasil perhitungan variabel dengan nilai $VIF < 15$ diperlihatkan pada Tabel 2 dibawah.

	variables	VIF
0	Usia	1.266009e+00
1	Sumatera	2.696329e+00
2	Senioritas	4.864156e+00
3	Trigliserida (mg/dL)	7.703643e+00
4	Visceral Fat	1.115201e+01

Tabel 2. Hasil Perhitungan dengan VIF<15

Variabel-variabel tersebut yang akan digunakan pada proses trainning model. Kemudian setelah manangani masalah multikolinearitas, kita perlu membuat beberapa model dengan kombinasi beberapa parameter/variabel dan melakukan uji kecocokan model (Goodness of Fit Test). Untuk pemilihan model sendiri berdasarkan prinsip parsimoni yaitu prinsip yang menyatakan bahwa semakin sederhana sebuah model statistik dengan jumlah variabel dependen cukup informatif untuk menjelaskan model, semakin baik pula model statistik tersebut. Disini metrik yang akan digunakan adalah AIC (Akaike's Information Criteria) dan BIC (Bayesian Information Criteria). Pemilihan

metrik AIC dan BIC sendiri karena metrik ini dapat menangkap kompleksitas serta performa model kedalam suatu nilai yang bisa dibandingkan antar model. Dengan rumus AIC dan BIC seperti dibawah,

$$AIC = 2k - 2\ln(L)$$

$$BIC = k\ln(n) - 2\ln(L)$$

dimana k adalah jumlah parameter, $\ln(L)$ adalah log-likelihood, dan n adalah ukuran sampel dalam kasus kita adalah 1339 obeservasi atau titik data. Variasi model yang akan dibuat sendiri adalah model dengan jumlah serta kombinasi variabel yang berbeda dengan variabel yang digunakan adalah variabel $VIF < 15$. Hasil lengkap uji kecocokan model terlihat pada Tabel 3.

	Variabel	Jumlah Parameter	Log-Likelihood	AIC	BIC
0	[Usia, Trigliserida (mg/dL), Visceral Fat, Sum...	6	-392.684805	797.369609	828.540733
1	[Usia, Trigliserida (mg/dL), Senioritas]	4	-395.975490	799.950980	820.731729
2	[Usia, Trigliserida (mg/dL), Sumatera]	4	-395.991279	799.982559	820.763308
3	[Usia, Trigliserida (mg/dL), Visceral Fat, Sen...	5	-395.861646	801.723292	827.699228
4	[Usia, Trigliserida (mg/dL), Visceral Fat, Sum...	5	-395.941703	801.883405	827.859342
5	[Usia, Trigliserida (mg/dL)]	3	-398.449504	802.899007	818.484569
6	[Usia, Trigliserida (mg/dL), Visceral Fat]	4	-398.423341	804.846681	825.627430
7	[Trigliserida (mg/dL)]	2	-404.776135	813.552270	823.942644
8	[Usia, Senioritas]	3	-420.696034	847.392069	862.977631
9	[Usia, Sumatera]	3	-422.903757	851.807513	867.393075
10	[Usia]	2	-425.368935	854.737869	865.128244
11	[Usia, Visceral Fat]	3	-424.395351	854.790703	870.376265
12	[Senioritas]	2	-426.137030	856.274059	866.664434
13	[Sumatera]	2	-434.005374	872.010748	882.401122
14	[Visceral Fat]	2	-434.143023	872.286046	882.676420

Tabel 3. Hasil Uji Kecocokan Model

Terlihat bahwa model dengan AIC paling kecil adalah model dengan variabel “Usia”, “Trigliserda (mg/dL)”, “Visceral Fat”, “Sumatera”, dan “Senioritas” sedangkan model dengan BIC paling kecil adalah model dengan variabel “Usia” dan “Trigliserida (mg/dL)”. Kita akan menggunakan kedua model ini untuk diuji performa modelnya.

Terakhir kita akan menguji performa model yang dikuantifikasi oleh suatu metrik tertentu. Jika kita hubungkan dengan konteks perusahaan, tentu perusahaan tidak mau salah memprediksi karyawannya yang seharusnya berkolesterol tinggi karena akan memberikan efek berupa penurunan produktivitas perusahaan nanti kedepannya. Adapun sebaliknya, perusahaan tentu tidak mau salah memprediksi karyawannya yang seharusnya normal menjadi kolesterol tinggi karena terdapat biaya pengobatan serta

fasilitas kesehatan yang seharusnhya tidak perlu. Oleh karena disini, kita mengasumsikan bahwa kedua jenis kesalahan tersebut memiliki bobot yang sama atau kerugian yang ditimbulkan sama. Oleh karena itu, metrik yang paling tepat digunakan adalah Accuracy. Accuracy sendiri merupakan proporsi antara prediksi yang benar terhadap jumlah total prediksi. Kemudian untuk trainning sendiri, kita akan menggunakan train test split sebesar 80:20 dan algoritma machine learning Logistic Regression. Pemilihan algoritma Logistic Regression didasarkan pada kemudahan interpretasi-nya dibandingkan dengan algoritma klasifikasi lain seperti Decision Tree ataupun Neural Network. Selain itu, sesuai Exploratory Data Analysis, dataset memang berisi banyak noise yang sudah coba diminimumkan dengan melakukan binning variabel dan sebagainya. Logistic Regression merupakan algoritma yang bisa dikatakan *simple* sehingga harapannya dapat *generalize* dengan baik ke data yang belum pernah dilihat pada proses trainning sebelumnya. Adapun hyparameter lengkap yang digunakan sesuai dengan Tabel 4.

Hyperparameter	Nilai
Train/Test Split	80/20
Algoritma	Logistic Regression
Metric	Accuracy

Tabel 4. Hyperparameter Trainning

2.5 HASIL PREDIKSI

Setelah model melalui proses trainning, model digunakan untuk memprediksi data pada testing set. Hasil prediksi model dapat dilihat pada Tabel 5,

Model	Variabel/Parameter	Accuracy
Logistic Regression (AIC minimum)	Usia, Trigliserida (mg/dL), Visceral Fat, Senioritas, Sumatera	89
Logistic Regression (BIC minimum)	Usia, Trigliserida (mg/dL)	89

Tabel 5. Hasil Prediksi Model

dimana ternyata kedua model memiliki performa yang sama. Hal ini menarik mengingat jumlah variabel atau parameter yang dipakai berbeda. Pada Gambar 17 diperlihatkan summary model pertama (AIC minimum) dengan koefisien serta konstanta masing-masing variabel hasil trainning dengan metode maximum likelihood estimation (MLE).

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                     CT   No. Observations:                 1066
Model:                          Logit   Df Residuals:                     1060
Method:                           MLE   Df Model:                            5
Date:                Tue, 16 Apr 2024   Pseudo R-squ.:                 0.09213
Time:                        14:52:16   Log-Likelihood:                -423.15
converged:                       True   LL-Null:                       -466.09
Covariance Type:            nonrobust   LLR p-value:                 4.907e-17
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                    -3.3757      0.319    -10.597      0.000      -4.000      -2.751
Usia                      0.9830      0.264      3.719      0.000       0.465       1.501
Trigliserida (mg/dL)      0.0118      0.002      6.242      0.000       0.008       0.016
Visceral Fat             -0.0137      0.037     -0.371      0.711      -0.086       0.059
Senioritas                0.1617      0.109      1.485      0.138      -0.052       0.375
Sumatera                  0.6933      0.317      2.188      0.029       0.072       1.314
============================================================================================
```

Gambar 17. Hasil Estimasi Koefisien

Terlihat bahwa variabel yang paling berpengaruh adalah “Usia” dengan koefisien sebesar 0,9830 disusul oleh “Sumatera”, “Senioritas”, “Visceral Fat”, dan “Trigliserida (mg/dL)”. Jika kita modelkan dalam bentuk matematis akan terlihat seperti berikut,

$$logit\big(p(x)\big) = \log\left(\frac{p(x)}{1-p(x)}\right) = -3{,}3757 + 0{,}9830 \times Usia + 0{,}6933 \times Sumatera +$$

$$0{,}1617 \times Senioritas - 0{,}0137 \times Visceral\ Fat + 0{,}118 \times Trigliserida\ (mg/dL)$$

BAB III
PENUTUP

3.1 KESIMPULAN

3.2 SARAN

Seseorang yang memiliki kolesterol tinggi berpotensi mengalami penyakit kardiovaskular disarankan untuk selalu menjaga kesehatannya

- **Jika diterapkan terhadap kasus nyata sebuah perusahaan**

Berdasarkan kesimpulan di atas, beberapa langkah preventif penanganan penyakit kardiovaskular yang dapat dilakukan oleh perusahaan adalah :

1. Pemberlakuan jam kerja yang efektif dan waktu istirahat yang cukup di sela jam kerja untuk karyawan perusahaan.
2. Pembuatan program kesehatan karyawan seperti *medical check-up*, vaksinasi, fasilitas kelas olahraga, layanan konseling kesehatan mental, pemberian boost vitamin secara rutin, dan penyediaan peralatan kantor yang mendukung active lifestyle.

3. Pembuatan SOP berkaitan dengan kesehatan karyawan. Yang mengatur bahwa karyawan yang memiliki atau berpotensi memiliki penyakit kardiovaskular akan diberikan potongan gaji untuk asuransi kesehatan dari karyawan tersebut.

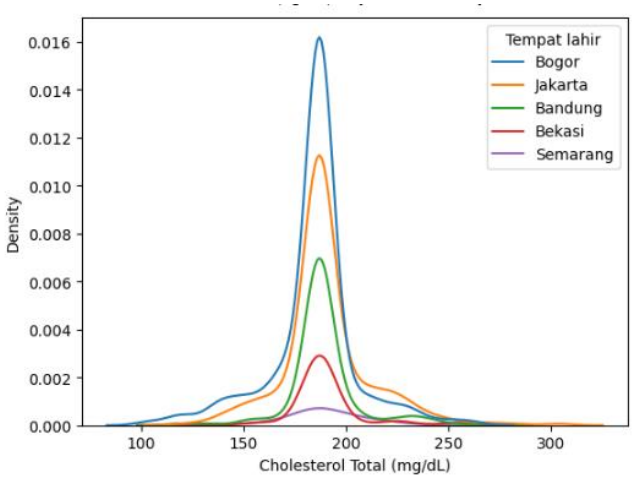
Langkah ini tentu lebih murah daripada perusahaan harus menanggung rugi karena efektivitas kinerja karyawan yang berkurang karena karyawan sakit, sehingga membuat perlambatan pada eksekusi program kerja dan perlambatan ekskalasi bisnis. Jika terdapat karyawan sakit, perusahaan harus mengeluarkan biaya ekstra semisal biaya bantuan medis. Selain itu, jika ada karyawan yang meninggal dunia, perusahaan juga harus memberikan santunan serta biaya perekrutan dan pelatihan kerja karyawan baru yang tergolong tinggi.

DAFTAR PUSTAKA

- Tim Medis Siloam Hospitals. (2024). Ketahui Kadar Kolesterol Normal Wanita & Pria Sesuai Usia. Diakses 10 April 2024, dari: https://siloamhospitals.com/informasi-siloam/artikel/kadar-gula-darah-puasa-normalhttps://siloamhospitals.com/informasi-siloam/artikel/kadar-gula-darah-puasa-normal
- Tim Medis Siloam Hospitals. (2024). Mengenal Kadar Gula Darah Puasa Normal dan Cara Menjaganya. Diakses 10 April 2024, dari: https://www.siloamhospitals.com/informasi-siloam/artikel/kadar-kolesterol-normalhttps://www.siloamhospitals.com/informasi-siloam/artikel/kadar-kolesterol-normal
- Cleveland Clinic. (2022). High Cholesterol: Causes, Symptoms and How It Affects the Body. Diakses 10 April 2024, dari: https://my.clevelandclinic.org/health/articles/11918-cholesterol-high-cholesterol-diseaseshttps://my.clevelandclinic.org/health/articles/11918-cholesterol-high-cholesterol-diseases
- Dr. Barbara Bowman. (2015). Heart Health is Good Business. Diakses 10 April 2024, dari: https://www.cdcfoundation.org/blog-entry/businesspulse-heart-health
- Center for Disease Control and Prevention. (2016). Worker Productivity: Cholesterol Evaluation Measures. Diakses 10 April 2024, dari: https://www.cdc.gov/workplacehealthpromotion/health-strategies/cholesterol/evaluation-measures/worker-productivity.html
- Center for Disease Control and Prevention. (2016). Heart Health is Good Business. Diakses 10 April 2024, dari: https://www.cdcfoundation.org/blog-entry/businesspulse-heart-health
- Heinsohn X Elerator. Seniority Level: Different Types Of Seniority At Work. Diakses 10 April 2024, dari: https://www.us.heinsohn.co/blog/seniority-level/https://www.us.heinsohn.co/blog/seniority-level/

- Kemenkes. (2018). Klasifikasi Obesitas setelah pengukuran IMT. Diakses 10 April 2024, dari: https://p2ptm.kemkes.go.id/infographic-p2ptm/obesitas/klasifikasi-obesitas-setelah-pengukuran-imt
- Henry Aidoo, Akye Essuman, Phyllis Aidoo, Anita O Yawson, and Alfred E Yawson. (2015). Health of the corporate worker: health risk assessment among staff of a corporate organization in Ghana. ,v.10;2015, 1-7. Doi:10.1186/s12995-015-0072-7. (Diakses 10 April 2024, dari: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4531895/)
- Healthy Working Lives. (2024). Supporting employees with cardiovascular disease. Diakses 10 April 2024, dari: https://www.healthyworkinglives.scot/workplace-guidance/ill-health-and-absence/supporting-employees-with-cardiovascular-disease/
- Tri Wahyuni, Jihanita Diansabila. (2020). Hubungan Indeks Massa Tubuh (IMT) dengan Kadar Kolesterol pada Mahasiswa Program Studi Kedokteran. Muhammadiyah Journal of Nutrition and Food Science, Vol 1, No 2 (2020). DOI: 10.24853/mjnf.1.2.54-59. Diakses 16 April 2024, dari: Hubungan Indeks Massa Tubuh (IMT) dengan Kadar Kolesterol pada Mahasiswa Program Studi Kedokteran | Wahyuni | Muhammadiyah Journal of Nutrition and Food Science (MJNF) (umj.ac.id)
- Yohana Margarita, Princen, Andi, Marcella Erwina Rumawas, Valentinus Budi Kidarsa, Bambang Sutrisna. (2013). Kadar Kolesterol Total dan Tekanan Darah Orang Dewasa Indonesia. Kesmas, Jurnal Kesehatan Masyarakat Nasional Vol. 8, No. 2. Diaksesn

LAMPIRAN



Gambar 1. Distribusi Cholesterol Total berdasarkan 5 tempat lahir mayoritas

Level Senioritas	Masa Kerja (tahun)
Junior	< 2
Associate	2-5
Mid-level	5-10
Senior	>10

Tabel 1. Level Senioritas berdasarkan perusahaan IT konsultan *Heinshohn Elevator*

Klasifikasi IMT	IMT (kg/m2)
Underweight	< 18.5
Normal	18.5-24.9
Overweight	25-29.9
Obesity	>30

Tabel 2. Klasifikasi IMT berdasarkan *World Health Organization (WHO)*

Klasifikasi Gula Puasa	Gula Puasa (mg/dL)
Normal	70-99
Prediabetes	100-125
Diabetes	>126

Tabel 3. Klasifikasi Gula Puasa berdasarkan *Siloam Hospitals*

Klasifikasi Trigliserida	Trigliserida (mg/dL)
Normal	<150
Tinggi	150-199
Waspada	>200

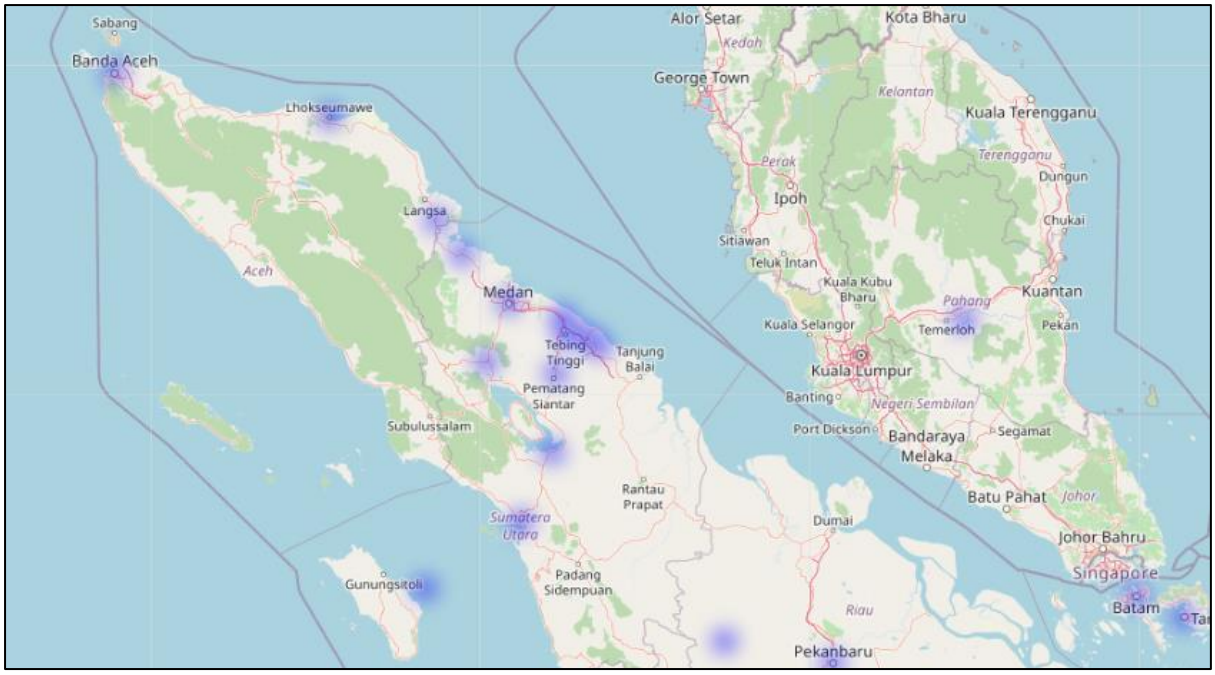
Tabel 4. Klasifikasi Trigliserida berdasarkan *Siloam Hospitals*

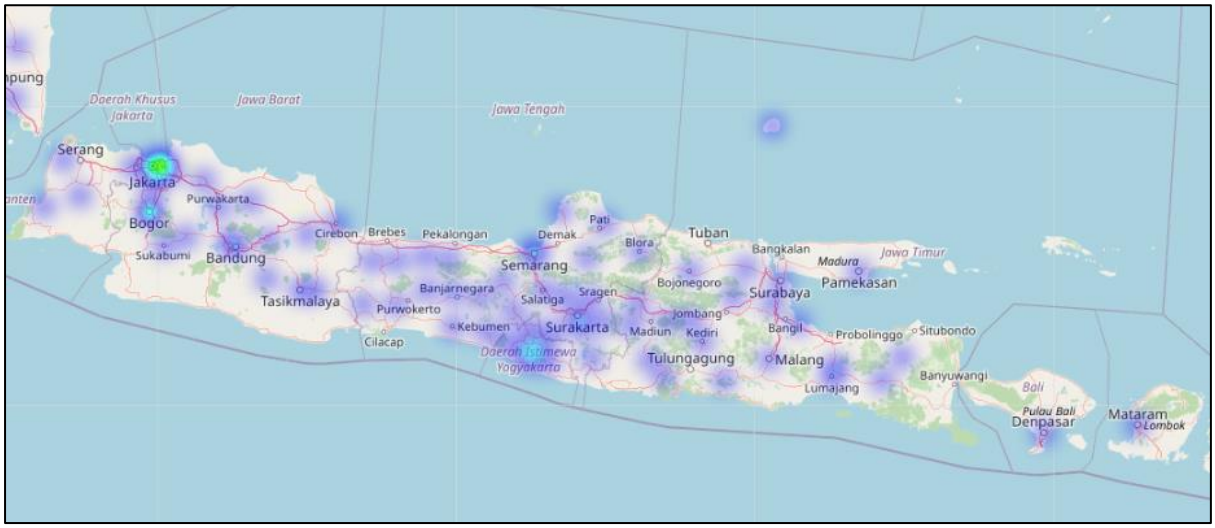
Klasifikasi Cholesterol Total	Cholesterol Total (mg/dL)
Normal	<200
Tinggi	>200

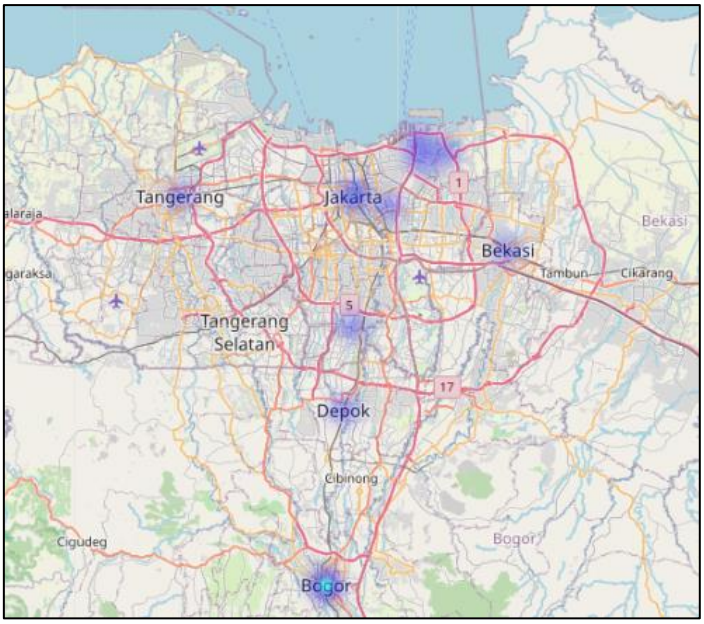
Tabel 5. Klasifikasi Cholesterol Total berdasarkan *Siloam Hospitals*

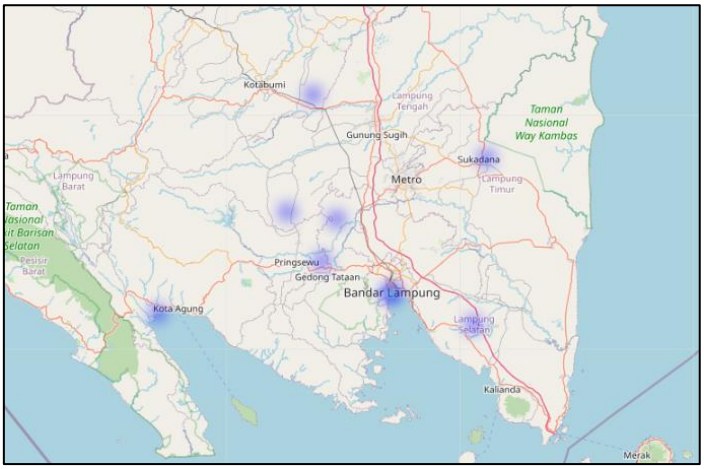
Rating TD	Tekanan Darah (S)	Tekanan Darah (D)
Normal	<120	<80
Elevated	120-129	<80
Stage 1	130-139	80-89
Stage 2	>140	>90

Tabel 6. Rating Tekanan Darah berdasarkan *Cedars Sinai*


Sabang
Banda Aceh
Lhokseumawe
Langsa
Aceh
Medan
Tebing Tinggi
Tanjung Balai
Pematang Siantar
Subulussalam
Gunungsitoli
Sumatera Utara
Rantau Prapat
Padang Sidempuan
Alor Setar
Kedah
George Town
Kota Bharu
Kelantan
Kuala Terengganu
Terengganu
Dungun
Perak
Ipoh
Chukai
Sitiawan
Teluk Intan
Kuala Kubu Bharu
Pahang
Kuantan
Kuala Selangor
Temerloh
Pekan
Kuala Lumpur
Banting
Negeri Sembilan
Port Dickson
Bandaraya Melaka
Segamat
Batu Pahat
Johor
Dumai
Johor Bahru
Singapore
Batam
Riau
Pekanbaru


Daerah Khusus Jakarta
Jawa Barat
Jawa Tengah
Serang
Jakarta
Purwakarta
Bogor
Sukabumi
Bandung
Cirebon
Brebes
Pekalongan
Demak
Pati
Blora
Tuban
Semarang
Bangkalan
Madura
Jawa Timur
Pamekasan
Surabaya
Bojonegoro
Banjarnegara
Salatiga
Sragen
Tasikmalaya
Purwokerto
Jombang
Kebumen
Surakarta
Madiun
Kediri
Bangil
Probolinggo
Situbondo
Cilacap
Daerah Istimewa Yogyakarta
Tulungagung
Malang
Lumajang
Banyuwangi
Bali
Pulau Bali
Denpasar
Mataram
Lombok


Tangerang
Jakarta
Bekasi
Tambun
Cikarang
Tangerang Selatan
Depok
Cibinong
Cigudeg
Bogor
Bogor
1
5
17


Kotabumi
Lampung Tengah
Taman Nasional Way Kambas
Gunung Sugih
Sukadana
Lampung Timur
Metro
Lampung Barat
Pringsewu
Gedong Tataan
Bandar Lampung
Kota Agung
Lampung Selatan
Kalianda
Merak