PREDIKSI PENYAKIT JANTUNG DENGAN MENGGUNAKAN *MACHINE LEARNING AUTOGLUON*

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Strata-1
Program Studi Teknik Industri - Fakultas Teknologi Industri
Universitas Islam Indonesia

Nama : Rafif Dzaki Muhammad
No. Mahasiswa : 19522027

PROGRAM STUDI TEKNIK INDUSTRI PROGRAM SARJANA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2024

PERNYATAAN KEASLIAN

PERNYATAAN KEASLIAN

Saya mengakui bahwa tugas akhir ini adalah hasil karya saya sendiri kecuali kutipan dan ringkasan yang seluruhnya sudah saya jelaskan sumbernya. Jika dikemudian hari ternyata terbukti pengakuan saya ini tidak benar dan melanggar peraturan yang sah maka saya bersedia ijazah yang telah saya terima ditarik kembali oleh Universitas Islam Indonesia.

Yogyakarta,06 Desember 2023

(Rafif Dzaki Muhammad)
19522027

SURAT BUKTI PENELITIAN

FAKULTAS
TEKNOLOGI INDUSTRI
Gedung KH. Mas Mansur
Kampus Terpadu Universitas Islam Indonesia
Jl. Kaliurang km 14,5 Yogyakarta 55584
T. (0274) 898444 ext. 4100, 4101
F. (0274) 895007
E. fti@uii.ac.id
W. fti.uii.ac.id

SURAT KETERANGAN PENELITIAN
Nomor: 264/Ka.Lab.Datmin/70/Lab.Datmin/XII/2023

Assalamu'alaikum Warahmatullahi Wabarakatuh

Kami yang bertanda tangan dibawah ini, menerangkan bahwa mahasiswa dengan keterangan sebagai berikut :

Nama	:	Rafif Dzaki Muhammad
No. Mhs	:	19522027
Dosen Pembimbing	:	Ir. Ira Promasanti Rachmadewi, M.Eng

Telah selesai melaksanakan penelitian yang berjudul " Prediksi Penyakit Jantung Dengan Menggunakan *Machine Learning AutoGluon*" di Laboratorium Data Mining, Program Studi Teknik Industri, Fakultas Teknologi Industri, Universitas Islam Indonesia tercatat mulai tanggal 14 Agustus 2023 sampai dengan tanggal 01 September 2023

Demikian surat keterangan kami keluarkan, agar dapat dipergunakan sebagaimana mestinya.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Yogyakarta, 07 Desember 2023

Kepala Laboratorium
Data Mining

Annisa Uswatun Khasanah, ST., M.B.A., M.Sc.

LEMBAR PENGESAHAN PEMBIMBING

LEMBAR PENGESAHAN PEMBIMBING

PREDIKSI PENYAKIT JANTUNG DENGAN MENGGUNAKAN *MACHINE LEARNING AUTOGLUON*

TUGAS AKHIR

Disusun Oleh :

Nama : Rafif Dzaki Muhammad
No. Mahasiswa : 19522027

Yogyakarta, 06 Desember 2023

Dosen Pembimbing

(Ir. Ira Promasanti Rachmadewi, M.Eng)

LEMBAR PENGESAHAN DOSEN PENGUJI

LEMBAR PENGESAHAN DOSEN PENGUJI

PREDIKSI PENYAKIT JANTUNG DENGAN MENGGUNAKAN *MACHINE LEARNING AUTOGLUON*

TUGAS AKHIR

Disusun Oleh :

Nama : Rafif Dzaki Muhammad

No. Mahasiswa : 19522027

Telah dipertahankan di depan sidang penguji sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata-1 Teknik Industri Fakultas Teknologi Industri Universitas Islam Indonesia

Yogyakarta, 22 Desember 2023

Tim Penguji

Ir. Ira Promasanti Rachmadewi, M.Eng
Ketua

Ir. Winda Nur Cahyo, S.T., M.T., Ph.D., IPM
Anggota I

Bambang Suratno, S.T., M.T., Ph.D
Anggota II

Mengetahui,
Ketua Program Studi Teknik Industri Program Sarjana
Fakultas Teknologi Industri
Universitas Islam Indonesia

Ir. Muhammad Ridwan Andi Purnomo, S.T., M.Sc., Ph.D., IPM
UII : 20101

HALAMAN PERSEMBAHAN

Alhamdulillahirabbil'alamin, Puji Syukur kehadirat Allah SWT yang telah memberikan rahmat dan hidayat-Nya sehingga penulis dapat menyelesaikan laporan Tugas Akhir ini. Tidak lupa penulis mengucapkan terima kasih dengan persembahan yang ditujukan kepada kedua orang tua penulis Bapak Dwi Taryono, S.Pd dan Ibu Tita Rosita, S.Pd yang telah memberikan semangat dan doa serta kakak penulis Syafiq Irfan Isnaindar, S.Kom yang telah memberikan masukan bagi penulis. Kemudian seluruh teman dan sahabat penulis yang telah memberikan dukungan dan dorongan untuk penulis.

MOTTO

Karena sesungguhnya sesudah kesulitan itu ada kemudahan

(QS Al-Insyirah ayat 5)

Barang siapa yang bersungguh-sungguh maka dia pasti berhasil

Jika anda tidak yakin bisa melakukannya maka anda tidak punya peluang sama sekali

(Arsene Wenger)

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabaraakaatuh

Alhamdulillahirabbil'alamiin¸segala puji dan syukur penulis sampaikan kepada Allah SWT atas berkah rahmat serta karunia-Nya sehingga penulisan dan penyusunan laporan Tugas Akhir dengan judul **"Prediksi Penyakit Jantung dengan Menggunakan *Machine Learning AutoGluon*"** yang bertujuan sebagai salah satu prasyarat mendapatkan gelar sarjana. Tidak lupa shalawat serta salam senantiasatercurahkan kepada junjungan kita Nabi Muhammad SAW yang telah membimbingseluruh manusia dari zaman kegelapan hingga zaman terang saat ini.

Selama penulisan Laporan Tugas Akhir, penulis ingin mengucapkan terima kasih kepada seluruh pihak yang membantu selama penulisan laporan berlangsung, baik secaralangsung maupun tidak langsung. Sehingga penulis ingin mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Ir. Hari Purnomo, M.T., IPU, ASEAN.Eng selaku Dekan Fakultas Teknologi Industri, Universitas Islam Indonesia.
2. Bapak Dr. Drs. Imam Djati Widodo. M.Eng.Sc. selaku Ketua Jurusan Teknik Industri Fakultas Teknologi Industri, Universitas Islam Indonesia.
3. Bapak Ir. Muhammad Ridwan Andi Purnomo, S.T., M.Sc., Ph.D., IPM selaku Ketua Program Studi Teknik Industri Program Sarjana, Fakultas Teknologi Industri, Universitas Islam Indonesia.
4. Ibu Ir. Ira Promasanti Rachmadewi, M.Eng. selaku dosen pembimbing laporan Tugas Akhir yang memberikan bimbingan, arahan, dan ilmu yang diberikan sehingga penulis dapat menyelesaikan laporan Tugas Akhir.
5. Kedua Orang tua penulis, Dwi Taryono, S.Pd. dan Tita Rosita, S.Pd. yang selalu memberikan doa, motivasi, dan dukungan moral maupun material kepada penulis serta kakak penulis Syafiq Irfan Isnaindar, S.Kom yang telah memberikan masukan kepada penulis.

6. Semua teman dan sahabat penulis, yang telah membantu penulis dalam penyelesaian laporan Tugas Akhir ini yang tidak dapat disebutkan satu per satu.

Penulis menyadari dalam penyusunan laporan Tugas Akhir ini masih terdapat kekurangan dan jauh dari sempurna, dengan besar hati penulis menerima kritik dan saran yang membangun. Penulis mengucapkan terima kasih kepada seluruh pihak yang membantu penulis dalam menyelesaikan laporan Tugas Akhir. Semoga kebaikan dan kemurahan hati oleh semua pihak yang membantu dalam penulisan laporan tugas akhir menjadi amal jariyah kebaikan dan bermanfaat bagi kita semua.

Yogyakarta, 1 September 2023

Penulis,

Rafif Dzaki Muhammad

ABSTRAK

Penyakit jantung merupakan penyakit yang termasuk ke dalam jenis penyakit kardiovaskular. Menurut *world health organization* (WHO) penyakit jantung khususnya di Indonesia menjadi penyakit dengan angka kematian tertinggi setelah penyakit stroke dengan jumlah kematian mencapai 85 untuk Wanita dan 107 untuk pria untuk 100.000 populasi. Angka kematian akibat penyakit jantung ini dapat terjadi akibat berbagai faktor seperti kurangnya aktivitas fisik, pola makan buruk, konsumsi rokok, dan alkohol. Kemudian juga terdapat faktor keadaan ekonomi dan kurangnya pengetahuan dalam merawat kesehatan jantung membuat abai terhadap kesehatan jantung. Pada penelitian ini dilakukan prediksi yang bertujuan untuk mengetahui hasil prediksi akurasi pada penyakit jantung. Selain itu juga untuk mengetahui model yang dapat memberikan hasil prediksi yang paling akurat. Metode penelitian yang digunakan dalam penelitian ini yaitu dengan menggunakan *machine learning AutoGluon*. *AutoGluon* sendiri merupakan *toolkit* yang dirancang oleh *Amazon Ltd* bersifat *open-source* yang dapat digunakan dengan mudah. *AutoGluon* memiliki layanan otomatisasi dalam melakukan pemrosesan data, seleksi model, arsitektur model, dan konfigurasi hyperparameter. Penelitian ini dilakukan menggunakan metode *supervised learning* dengan fokus terhadap model klasifikasi. Berdasarkan hasil penelitian yang telah dilakukan dapat diketahui bahwa model prediksi terbaik terdapat pada model *weigthed ensembled learning* yang memiliki nilai valensi terbasar dibandingkan model yang lain yaitu dengan nilai sebesar 0.939724. Kemudian nilai akurasi yang dihasilkan dalam penelitian ini yaitu sebesar 0.94.

Kata Kunci: Penyakit Jantung, *Machine Learning, AutoGluon*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

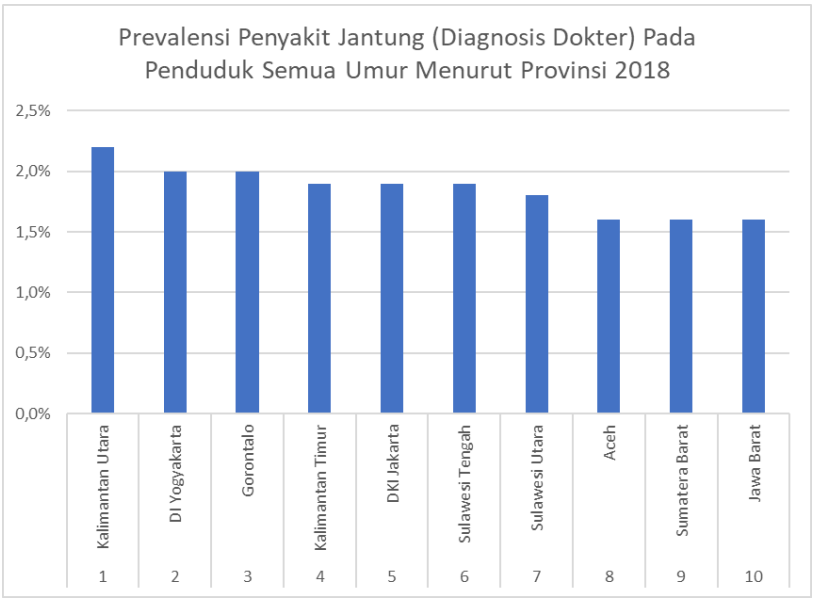
1.1 Latar Belakang

Penyakit non-infeksi atau yang lebih dikenal dengan penyakit tidak menular merupakan jenis penyakit yang berasal dari infeksi mikroorganisme seperti protozoa, bakteri, jamur, dan virus. Penyakit ini mengalami peningkatan kasus pada tiap tahun akibat lemahnya dalam pengendalian faktor risiko. Walaupun jenis penyakit ini tidak menular secara langsung, baik dari manusia ke manusia maupun dari hewan ke manusia. Akan tetapi penyakit tidak menular turut menyumbang angka kematian di dunia sedikitnya 70% kematian (Kementerian Kesehatan Republik Indonesia, 2020). Sebanyak 17 juta orang meninggal dunia akibat kematian dini pada penyakit tidak menular dengan 38% disebabkan oleh cardiovascular (World Health Organization, 2021).

Cardiovascular Diseases atau yang lebih dikenal sebagai penyakit kardiovaskular adalah penyakit yang menyerang bagian sekolompok gangguan jantung dan pembuluh darah (World Health Organization, 2021). Kardiovaskular dapat dihubungkan dengan kerusakan pada bagian arteri seperti jantung, mata, ginjal, maupun otak. Penyakit ini berhubungan dengan adanya penumpukan lemak pada arteri (*arteosklerosis*) dan meningkatkan risiko terjadinya pembekuan darah (NHS UK, 2022). Menurut (World Health Organization, 2021) penyakit kardiovaskular menjadi penyebab utama kematian di seluruh dunia dengan angka sebesar 17,9 juta orang pada tahun 2019. Angka tersebut mewakili 32% dari total keseluruhan kematian di seluruh dunia. Angka kematian akibat penyakit jantung diperkirakan akan terus meningkat hingga mencapai 23,6 juta angka kematian pada tahun 2030 (Aha Statictical Update, 2018).

Ishemic heart disease atau penyakit jantung koroner menjadi salah satu penyumbang angka kematian di Indonesia setelah stroke. Jumlah kematian tersebut mencapai 85 kematian untuk Wanita dan 107 kematian untuk pria per 100.000 populasi pada tahun 2019 (World Health Organization, 2019). Kalimatan Utara menjadi provinsi

di Indonesia dengan prevalensi penyakit jantung tertinggi menurut diagnosis dokter, yang kemudian disusul oleh Daerah Istimewa Yogyakarta dan Gorontalo. Berikut ini merupakan grafik 10 besarprovinsi di Indonesia dengan prevalensi penyakit jantung tertinggi menurut diagnosis dokter pada tahun 2018 (Hasil Utama Riskesdas 2018, 2018).



Gambar 1. 1 Grafik Prevalensi Penyakit Jantung Tahun 2018

Sumber : Riset Kesehatan Dasar 2018

Penyakit kardiovaskular dapat terjadi karena berbagai faktor seperti pola makan yang buruk, kurangnya aktivitas yang melibatkan fisik, merokok dan mengkonsumsi alkohol yang berbahaya. Dampak yang dapat terjadi akibat faktor risiko tersebut berupa hipertensi, hiperlimidea, dan obesitas. Sehingga berisiko dalam meningkatkan terserang penyakit jantung, stroke, gagal jantung, maupun penyakit kompilasi lainnya. Terdapat beberapa langkah yang dapat dilakukan untuk mengurangi risiko terjadinya penyakit kardiovaskular seperti mengurangi konsumsi garam dalam makanan, memperbanyak konsumsi sayuran dan buah-buahan, berhenti merokok serta menjauhi konsumsi alkohol yang berbahaya (World Health Organization, 2021).

Selain melakukan langkah-langkah yang dilakukan untuk mengantisipasi risiko penyakit kardiovaskular, memeriksa kesehatan secara rutin juga penting dalam menjaga

kesehatan jantung. Akan tetapi faktor keadaan ekonomi dan kurangnya pengetahuan mengenai merawat kesehatan jantung bagi masyarakat menengah kebawah membuat masyarakat abai terhadap kesehatan jantung terlebih biaya pemeriksaan kesehatan rutin yang mahal (Lestari, 2022). Hal ini diperkuat pada jumlah kasus dan biaya katastropik yang tinggi pada tahun 2021 sebagai berikut (Profil Kesehatan Indonesia 2021).

Tabel 1. 1 Jumlah Kasus dan Biaya Katastropik Program JKN Tahun 2021
Sumber : Profil Kesehatan Indonesia Tahun 2021

No	Katastx'ropik	Realisasi s/d 31 Desember 2021		
		Kasus	Biaya	Rata-rata Biaya per kasus
1	Jantung	12.934.931	8.671.706.289.303	670.410
2	Kanker	2.595.520	3.500.655.437.003	1.348.730
3	Strok	1.992.014	2.163.344.987.900	1.086.009
4	Gagal Ginjal	1.417.104	1.781.134.745.860	1.256.884
5	Thalasemia	281.577	604.616.997.602	2.147.253
6	Hemofilia	98.225	590.659.296.753	6.013.330
7	Leukimia	137.749	364.611.205.552	2.646.925
8	Sirosis Hepatis	160.152	238.497.880.605	1.489.197
Total		**19.617.272**	**17.915.226.840.578**	

Dalam mengatasi permasalahan diatas, media alternatif yang dapat membantu memprediksi penyakit jantung dapat menjadi salah satu solusi dalam menyelesaikan permasalahan diatas melalui peningkatan model prediksi (Apriyanto Alhamad, 2019). Perkembangan teknologi yang berkembang pesat memungkinkan untuk diterapkan dalam bidang ilmu kedokteran dengan memanfaatkan teknologi *machine learning* dan *deep learning* yang mampu menganalisis volume data dengan jumlah yang besar maupun mendeteksi pola serta tren data. Penerapan teknologi ini diharapkan dapat meningkatkan kualitas perawatan kesehatan, khususnya dalam menginisiasi deteksi dan prognosis penyakit dengan lebih akurat (Fayeza Sifat Fatima, 2022). Otomatisasi dalam memprediksi suatu penyakit dapat menciptakan *platform* dengan data terstruktur yang dapat memberikan perawatan yang efektif bagi pasien. Sehingga dapat mengubah personalisasi dalam tingkat perawatan kesehatan dengan penerapan *Artificial Intellegence* dan *Machine Learning*. Selanjutnya komputer mempelajari jenis pola penyakit yang dialami dan mengubahnya menjadi data terstruktur untuk memprediksi hal tersebut (M. Swathy, 2022).

Penelitian ini menggunakan *machine learning* sebagai *tools* dalam melakukan prediksi penyakit dibadingkan dengan *operation research*. *Machine learning* merupakan metode yang digunakan untuk menghasilkan model matematis yang menggambarkan berbagai pola data (Putra, 2020). *Machine learning* membantu agar pekerjaan menjadi lebih mudah serta dapat menyelesaikan permasalahan (Fangatulo Dodo, 2019). *Machine learning* memiliki keterikatan dengan pertanyaan mengenai bagaimana menciptakan program komputer yang mampu berkembang secara otomatis melalui pengalaman (Mitchell, 1997).

Sedangkan *operational research* (Inggris) atau *operations research* (Amerika Serikat) merupakan metode ilmiah yang diterapkan dalam penggunaan sumber daya dengan optimal dan efisien untuk memecahkan suatu permasalahan yang muncul ketika melakukan tugas (Murdifin Haming, 2019). Pemilihan *machine learning* dibandingkan *operations research* dalam melakukan penelitian ini karena *machine learning* mampu mengetahui cara mengatasi data dalam jumlah besar yang dapat melakukan pemodelan secara kompleks. Kemudian *machine learning* dapat mengatasi permasalahan hanya melalui data tanpa menetapkan konsep yang telah ditetapkan sebelumnya (Alvaro Talavera, 2020).

Penelitian ini dilakukan dengan menggunakan *machine learning* dengan metode *supervised learning*, dimana *supervised learning* biasa dikenal dengan model prediktif. Model ini bekerja dengan data historis yang diberikan arahan yang mengenai hal yang dipelajari (Mohamad Adhisyanda Aditya, 2020). *Supervised learning* merupakan model *machine learning* dengan nilai variabel yang terikat dimana sebelumnya nilai pada variabel tersebut telah ketahui (Nur Baiti Nasution, 2023). *Supervised learning* bekerja dengan *data training* yang telah diklasifikasikan secara detail dan digunakan untuk uji coba pada *data testing*, sehingga mendapatkan hasil sesuai harapan pada *data training* (Fathurohman, 2021).

Supervised learning yang digunakan dalam penelitian ini yaitu klasifikasi. Klasifikasi adalah proses yang dilakukan daalam mengevaluasi objek data yang nantinya akan dimasukkan ke dalam kategori kelas tertentu berdasarkan kelas yang ada (Dito Putro Utomo, 2020). Model klasifikasi pada metode *supervised learning* digunakan untuk

melakukan prediksi pada nilai atribut target kategorikal dimana model klasifikasi dapat menghasilkan aturan yang memungkinkan dalam melakukan prediksi pada kelas target dengan contoh masa depan. Model klasifikasi dilakukan dengan langkah awal untuk melakukan observasi di masa lampau pada kelas target yang telah diketahui (Vercellis, 2009). Model klasifikasi kerentanan pada penyakit jantung dapat digunakan dalam mendeteksi dini penyakit jantung dengan menggunakan *machine learning* dalam model klasifikasi berdasarkan kinerja pada proses dan hasil penilaian (Wiji Lestari, 2023). Dengan memanfaatkan data rekam medis pasien yang mengidap penyakit kardiovaskular untuk membuat model prediksi dirisiko dini, data tersebut dapat diolah dengan menggunakkan *machine learning* untuk membuat klasifikasi risiko kematian akibat penyakit kardiovaskular (Ahmadien Hafizh Yusufi, 2022). Klasifikasi adalah bagian keilmuan dalam *machine learning* yang mampu menangani *big data*. Metode yang terdapat dalam *supervised learning* dalam melakukan klasifikasi seperti *Logistic Regression, K-Nearest Neighboor, Support Vector Machine, Naïve Bayes, Decision Tree,* dan *Random Forest* (Fajar Sodik Pamungkas, 2020).

Logistic regression merupakan algoritma yang menghubungkan variabel independen terhadap variabel terikat dengan bentuk kategori dengan nilai kategori 0 dan 1 untuk melakukan prediksi dengan jenis regresi dalam menghitung probabilitas (Manzilur Rahman Romadhon, 2021). Kemudian *K-Nearest Neighbors* (KNN) merupakan algoritma yang digunakan dalam mencari nilai pada kelompok k data training yang dekat dengan objek pada data pengujian dalam melakukan klasifikasi (Dewi Cahyantia, 2020). Selanjutnya *Support vector machine* adalah algoritma yang biasa digunakan dalam melakukan klasifikasi pada data biner dengan tujuan untuk melakukan identifikasi terhadap *hyperlane* yang dilakukan secara efektif (Zhongming Wu, 2023). *Naïve Bayes* merupakan model algoritma klasifikasi yang dilakukan dengan probabilistik secara sederhana berdasarkan penerapan teorema bayes terhadap independensi (ketidaktergantungan) yang kuat (Taghsya Izmi Andini, 2016). *Decision Tree* merupakan algoritma yang digunakan untuk melakukan klasifikasi sampel data yang kelasnya belum diketahui kedalam kelas yang tersedia. *Decision Tree* berbentuk seperti struktur pohon dimana internal node menunjukkan pengujian suatu atribut, cabang menunjukkan output

pengujian, dan leaf node menunjukkan distribusi kelas (Laila Qadrini, 2021). *Random Forest* adalah pengembangan dari *decision tree* yang bekerja melalui pemilihan atribut yang diacak pada setiap node dalam melakukan klasifikasi, dimana proses klasifikasi dilakukan berdasarkan pohon keputusan yang dikembalikan dalam penerimaan suara terbanyak (Luthfiana Ratnawati, 2019).

Hal ini berbeda dengan metode *unsupervised learning* maupun *reinforcement learning*. *Unsupervised learning* atau dikenal dengan metode deskriptif dimana dalam metode ini tidak terdapat target yang ditetapkan dan faktor pendukung lainnya. Contoh dari metode *machine learning* ini adalah *K-Means Clustering* (Mohamad Adhisyanda Aditya, 2020). Tujuan dari *unsupervised learning* untuk mengidentifikasi kelompok dengan catatan homogen atau bersifat sama yang dikenal dengan *cluster* (Vercellis, 2009). Sedangkan *Reinforcement learning* bertujuan untuk melakukan efisiensi kinerja secara maksimal dimana mesin melakukan *training* untuk membuat keputusan khusus yang didasarkan pada kebutuhan. Mesin yang melakukan *training* secara terus menerus dapat melakukan pemecahan masalah dengan penguatan konsep dengan lingkungan sekitarnya. Contoh dari metode *reinforcement learning* adalah *Markov Decision Process* (Mohamad Adhisyanda Aditya, 2020). *Reinforcement learning* menerapkan sistem pengambilan keputusan dengan *trial and error* dengan mengeksplorasi lingkungan yang tidak pasti dan tidak diketahui sebelumnya dalam menggapai suatu tujuan (Oluwaseyi Ogunfowora, 2023).

Prediksi yang dilakukan dalam penelitian ini dilakukan dengan menggunakan *AutoGluon*. *AutoGluon* merupakan *toolkit* dalam *Machine Learning* bersifat *open-source* yang dirancang oleh Amazon Ltd yang mudah digunakan serta dapat diperluas. *AutoGluon* memiliki layanan otomatisasi dalam pemrosesan data, seleksi model, arsitektur model, dan konfigurasi hypermeter (Wenwen Qi, 2021). *AutoGluon* merupakan *open source* dalam kerangka kerja *AutoML* dalam rangkaian Python sebagai model *training Machine Learning* pada data tabular yang belum diproses. *AutoGluon Tabular* dapat menyatukan beberapa modeldan menyusunnya dalam beberapa lapisan (Sanjiv R. Da, 2022). *AutoGluon Tabular* adalah salah satu *library* dalam *Python* yang kompatibel digunakan dalam berbagai macam model, mudah digunakan, serta akurat dengan *AutoML*

terhadap data tabular. *AutoGluon Tabular* dapat melakukan *processing data* tingkat lanjut, *deep learning*, maupun *multi-layer ensembling* yang dapat mengenali tipe data di setiap kolom untuk melakukan *pre-processing data* secara otomatis (Nick Erickson, 2020).

Dibandingkan dengan model *machine learning* non-*autogluon*, *AutoGluon Tabular* dengan dapat mudah digunakan dengan penyajian API (*Application Programming Interface*) melalui beberapa prinsip. Prinsip sederhana dimana *user* dapat melakukan *training* model dengan hanya beberapa baris kode. Prinsip ketangguhan dengan *user* mempersiapkan data mentah tanpa melakukan manipulasi data maupun merekayasa fitur. Prinsip waktu yang dapat diprediksi oleh *user* dengan melakukan pembatasan waktu dan menemukan model terbaik yang dilakukan oleh *user*. Prinsip toleransi pada kesalahan dimana *user* ketika mengalami gangguan dapat melanjutkan *training* dengan memeriksa seluruh langkah sebelumnya (Prasanna, 2020).

Selain itu jika dibandingkan dengan *machine learning* non-*autogluon*, *AutoGluon* sendiri memiliki *Multi-Layer Ensembling* dimana pada lapisan pertama memiliki beberapa model dasar dengan *output* yang digabung dan dimasukkan ke lapisan berikutnya. Dimana lapisan pertama dengan nilai *hyperparameter* yang sama nantinya digunakan kembali sebagai alternatif dalam *deep learning* yang memanfaatkan latihan berlapis yang menghubungkan antar lapisan. Model dengan antar lapisan ini tidak hanya mengambil *input* dari model prediksi pada lapisan sebelumnya, tetapi dapat memanfaatkan *feature* yang ada pada data tersebut (Nick Erickson, 2020). Sehingga penelitian ini dapat dilakukan dengan menggunakan *AutoGluon*. *AutoGluon* sendiri dapat menyimpulkan jenis masalah prediksi berdasarkan nilai yang terdapat pada kolom label apabila *user* tidak menentukan sebelumnya. Selain itu Optimasi dalam AutoGluon dapat membantu *user* dalam menerjemahkan *raw data* menjadi prediksi yang akurat dengan cepat (Nick Erickson, 2020).

Dataset yang digunakan dalam penelitian diunduh melalui situs *Kaggle.com* yang berasal dari CDC (*Centers forDisease Control and Prevention*) yang di survey dalam kegiatan BRFSS (*Behavioral Risk Factor Surveilance System*) di Amerika Serikat. Pengunduhan dataset dari *Kaggle.com* dalam penelitian ini karena keterbatasan dalam

mengakses dataset secara langsung di rumah sakit yang bersifat privasi. Sehingga dengan penelitian ini diharapkan dapat menjadi referensi dalam melakukan prediksi penyakit, khususnya dengan menggunakan *AutoGluon*.

1.2 Rumusan Masalah

Berikut rumusan masalah pada penelitian ini sebagai berikut:

1. Bagaimana hasil prediksi penyakit Jantung dengan menggunakan *AutoGluon*?
2. Apa model yang memberikan hasil prediksi yang paling akurat?

1.3 Tujuan Penelitian

Berikut merupakan tujuan dilakukannya penelitian ini sebagai berikut:

1. Untuk mengetahui hasil prediksi penyakit Jantung dengan menggunakan *AutoGluon*.
2. Untuk mengetahui model yang memberikan hasil prediksi yang paling akurat.

1.4 Manfaat Penelitian

Berikut merupakan manfaat pada penelitian yang dilakukan sebagai berikut:

1. Penelitian yang dilakukan dengan *AutoGluon* masih jarang dilakukan, karena *AutoGluon* baru diriilis oleh *Amazon Web Services* (AWS) pada tahun 2019. Sehingga penelitian ini diharapkan dapat membantu peneliti lain sebagai referensi dalam penelitian selanjutnya, baik dalam bidang Kesehatan maupun dalam bidang teknologi yang lainnya.
2. Mengetahui hasil prediksi penyakit jantung berdasarkan dataset *Cardiovascular Diseases Risk Prediction*.
3. Mengetahui metode yang dapat memberikan akurasi hasil prediksi paling terbaik.

1.5 Batasan Penelitian

Berikut merupakan batasan masalah yang diterapkan pada penelitian ini yaitu.

1. Penelitian ini dilakukan dengan jenis *machine learning* dengan jenis *supervised learning* yang dilakukan dalam mencari hasil klasifikasi.

2. Penelitian ini berfokus pada penerapan *AutoGluon* untuk mendapatkan hasil prediksi penyakit jantung.
3. Dataset yang digunakan dalam penelitian ini yaitu dataset dari CDC (*Centers for Disease Control and Prevention*) yang di survey dalam kegiatan BRFSS (*Behavioral Risk Factor Surveilance System*) pada tahun 2021 yang dapat diakses melalui situs Kaggle.com.

BAB II

TINJAUAN PUSTAKA

2.1 Landasan Teori

Landasan teori merupakan teori yang digunakan dalam mendukung penulis dalam membantu selama proses penelitian.

2.1.1 Kardiovaskular

Cardiovascular Disease (CVDs) atau yang dikenal dengan penyakit Kardiovaskular merupakan sekelompok penyakit yang mempengaruhi jantung maupun pembuluh darah yang disebabkan oleh berbagai faktor, seperti sosial ekonomi, perilaku, lingkungan hingga konsumsi alkohol maupun rokok. Selain itu juga terdapat pengaruh risiko berdasarkan Riwayat keluarga seperti etnis, usia, dan jenis kelamin (World Heart Federation, 2023). Penyakit Kardiovaskular dapat memungkinkan sesorang dapat merasakan *Symtomatic* (sakit yang dapat dirasakan secara fisik) maupun *Asymptomatic* (sakit yang tidak merasakan apa-apa). Penyakit kardiovaskular termasuk kedalam masalah jantung dan pembuluh darah seperti penyempitan pada jantung maupun organ lain, masalah jantung dan pembuluh darah ketika lahir, katup jantung tidak berfungsi dengan baik, serta detak jantung tidak teratur (Cleveland Clinic, 2022). Terdapat 4 jenis utama yang umum *cardiovascular disease* (NHS UK, 2022):

1. Penyakit jantung koroner

   Penyakit jantung koroner terjadi saat suplai darah otot jantung terganggu akibat timbunan lemak (atheroma) pada arteri koroner. Penyakit ini dapat mengakibatkan angina (nyari dada) jika arteri koroner mengalami penyempitn akibat pembentukan atheroma dan pasokan darah menuju otot jantung terbatas. Sehingga apabila arteri korener tersumbat dapat mengakibatkan serangan jantung.

2. Stroke

   Stroke terjadi ketika suplai darah menuju otak terganggu, dimana otak membutuhkan persediaan oksigen dan nutrisi yang stabil supaya dapat berfungsi dengan baik. Sehingga apabila aliran darah terhenti dapat mengkibatkan sel-sel otak yang mulai mati dan berakibat kerusakan otak yang dapat menyebabkan kematian.

3. Penyakit arteri perifer

   Penyakit arteri perifer terjadi jika terdapat penyumbatan pada arteri menuju anggota tubuh. Gejala penyakit ini pada umumnya berupa rasa nyeri pada bagian kaki ketika berjalan. Jika keadaan semakin memburuk, kaki akan merasa berat saat berjalan ketika digunakan untuk berlatih.

4. Penyakit Aorta

   Penyakit aorta yang umumnya terjadi yaitu dengan jenis aneurisma aorta. Penderita penyakit ini biasanya akan merasakan rasa nyaeri pada dada, punggung, maupun perut yang disebabkan oleh diding aorta yang membengkak dan menonjol.

2.1.2 Penyakit Jantung

Penyakit jantung merupakan penyakit yang umumnya merujuk pada masalah yang dapat memberikan dampak bagi kinerja jantung. Penyakit jantung berbeda dengan penyakit kardiovaskular, dimana organ tubuh yang mengalami masalah pada penyakit jantung terdapat di bagian hati. Sedangkan pada penyakit kardiovaskular organ tubuh yang mengalami masalah terdapat pada organ tubuh pembuluh darah dan jantung (Annisa, 2019). Penyakit jantung umunnya terjadi akibat kurangnya oksigen yang diedarkan oleh darah menuju pembuluh darah di jantung, sehingga menyebabkan rusaknya sel yang terdapat pada otot-otot jantung yang digunakan untuk mengalirkan aliran darah menuju seluruh tubuh. Selain itu juga penyakit jantung dapat mengakibatkan organ jantung gagal dalam memompa darah yang disebabkan kejantung yang terjadi di otot jantung (Eka Wahyudi, 2017). Terdapat faktor gaya hidup maupun kondisi medis yang dapat mempengaruhi risiko penyakit jantung yaitu (CDC, 2023):

1. Obesitas
2. Diabetes
3. Konsumsi alkohol berlebihan
4. Kurangnya aktivitas fisik
5. Kebiasaan makan tidak sehat

Kemudian menurut (Dewi, 2021) faktor dalam penyakit jantung terdiri dari dua faktor yaitu faktor yang dapat diubah dan faktor yang tidak dapat diubah. Faktor yang dapat diubah terdiri dari obesitas, konsumsi alkohol yang berlebihan, kurangnya aktivitas fisik, diabetes, kolestrol tinggi, dan diabetes. Sedangkan faktor yang tidak dapat diubah yaitu keturunan, jenis kelamin, dan usia.

Body Mass Index (BMI) atau Indeks Masa Tubuh (IMT) digunakan untuk menilai rasio tinggi badan dengan berat badan dalam memprediksi jumlah lemak dalam tubuh. BMI dihitung dengan membagi antara berat badan (kg) dengan kuadrat tinggi badan (m^2). Jumlah lemak tubuh yang berlebihan dapat mengakibatkan berbagai penyakit seperti penyakit jantung, diabetes tipe 2, dan stroke. Selain itu apabila jumlah lemak tubuh yang kurang dapat menyebabkan malnutrisi. Sehingga jumlah lemak tubuh harus sesuai untuk mendukung mineral dan vitamin masuk ke dalam tubuh yang menyajikan energi bagi tubuh untuk melindungi organ dan suhu tubuh (Cleveland Clinic, 2022). BMI ideal bagi orang dewasa umumnya berada di rentang skala 18,5 hingga 24,9. Sedangkan BMI ideal bagi anak-anak hingga remaja dengan usia 2 sampai 18 tahun dapat dinilai berdasarkan usia dan kondisi fisik berupa jenis kelamin, tinggi badan, dan berat badan. Berikut merupakan nilai rentang skala dalam BMI (NHS UK, 2022):

Tabel 2. 1 Rentang Skala BMI

Skala	Keterangan
< 18,5	Kondisi kurang berat badan
18,5 – 24,9	Kondisi berat badan sehat
25 – 29,9	Kondisi kelebihan berat badan
> 30	Kondisi obesitas

Arthritis merupakan penyakit yang terjadi pada sendi atau tempat bertemunya antar tulang yang mengakibatkan kerusakan. Apabila usia bertambah dapat mempengaruhi

kondisi persendian yang menyebabkan melemah secara alami. Akan tetapi radang sendi dapat terjadi dalam kondisi kesehatan tertentu (Cleveland Clinic, 2023).

2.1.3 *Machine Learning*

Machine Learning merupakan salah satu jenis *Artificial Intelligence* (AI) yang memungkinkan komputer untuk mempelajari data tanpa harus mengikuti program yang telah diinstruksikan dengan fokus pengembangan program baru komputer untuk dapat berkembang ketika terdapat data baru. *Machine Learning* adalah program komputer yang dapat bekerja dengan mengoptimalkan kinerja melalui penggunaan data sampel atau pengalaman masa lalu (Budiharto, 2016). *Machine Learning* merupakan *Artificial Intelligence* dengan kemampuan untuk mempelajari mesin secara otomatis melalui data dan pengalaman dengan mengidentifikasi pola untuk membuat model prediksi. Teknologi dalam *machine learning* memungkinkan untuk memperoleh informasi secara detail melalui volume data yang memanfaatkan algoritma dalam mengidentifikasi pola serta pembelajaran yang dilakukan secara berulang. Algoritma dalam *machine learning* memanfaatkan metode komputasi untuk mempelajari data secara langsung melalui data dibandingkan dengan model persamaan yang telah ditentukan sebelumnya (Kanade, 2022).

Berikut merupakan jenis tipe-tipe dari machine learning:

1. *Supervised Learning*

   Supervised learning memanfaatkan informasi *input* dan *output* yang diinginkan melalui kumpulan *dataset training*. Sistem akan mempelajari pola data tersebut yang nantinya hasil pola tersebut akan menjadi referensi pada kumpulan data selanjutnya (Puput Santoso, 2021).

2. *Unsupervised Learning*

   Unsupervised learning bersifat deskriptif dimana biasa digunakan dalam mengkalsifikasikan data. Jenis *machine learning* ini memerlukan pembelajaran data yang telah ada sebelumnya karena tidak menerima *training dataset* yang bersifat prediktif (Puput Santoso, 2021).

3. *Reinforcement Learning*

   Reinforcement learning bekerja melalui informasi yang telah diterima sebelumnya berdasarkan tindakan yang telah dilakukan pada sebelumnya (Leonita Angelina, 2016). *Reinforcement learning* dilakukan dengan melakukan pengolahan data secara berulang-ulang dalam pembelajaran *trial and error* untuk menemukan hasil yang terbaik (Kolondam, 2021).

2.1.4 *Automated Machine Learning*

AutoML didefinisikan sebagai kombinasi antara otomatisasi dengan *machine learning*, dimana *AutoML* melibatkan *pipeline machine learning* yang dikerjakan secara otomatis dengan anggaran komputasi yang terbatas (Bahrynovska, 2022). *Automated Machine Learning* (*AutoML*) merupakan sebuah teknologi dalam *machine learning* yang membantu dalam pemolihan algoritma secara otomatis, mengoptimasi *hypermeter*, perulangan model, dan evaluasi model. AutoML berupaya dalam membantu *data scientist* dalam tugas yang bernilai tinggi dan meningkatkan nilai akurasi model *machine learning* (Nick Erickson, 2020).

Berikut merupakan *package* yang telah dikembangkan oleh *AutoML* yaitu (AutoML.org, 2023) :

1. AutoWEKA

   Pendekatan dalam pemilihan algoritma *machine learning* dan *hpermater* yang dilakukann secara simulatan dan dikombinasikan dengan paket WEKA untuk menghasilkan model yang baik untuk berbagai jenis data.

2. Auto-sklearn

   Perluasan dari AutoWEKA yang menggantikan klasifikasi dan regressir scikit-learn biasa dengan menggunakan *library scikit-learn*.

3. Auto-Pytorch

   Kerangka *deep learning* PyTorch yang mengoptimalkan hypermeter dan arsitektur neural.

4. AutoGluon

   Pendekatan yang dilakukan dengan penumpukan pada beberapa lapisan pada model *machine learning* yang berbeda.

5. H2O AutoML

   Menyediakan pemodelan dan sistem seleksi secara otomatis dalam *machine learning* H2O dan *platform* analisis data.

6. MLBoX

   Library AutoML yang terdiri dari *pre-processing*, optimalisasi, dan prediksi.

7. TPOT

   Sebuah asistem yang digunakan dalam *data sciene* untuk mengoptimalkan alur kerja *machine learning* dengan pemrograman genetik.

8. TransmogrifAI

   Library dalam *AutoML* yang bekerja diatas *Spark*

2.1.5 *AutoGluon*

AutoGluon merupakan *open source* dari AWS *(Amazon Web Series)* yang dapat diakses dengan mudah oleh siapa saja untuk membantu *machine learning* dan *deep learning* dalam menentukan model yang tepat pada tugas tertentu (Levande, 2021). *AutoGluon* melakukan otomatisasi dengan memanfaatkan sumber daya komputasi yang ada dalam menemukan model terbaik dengan waktu proses yang telah ditentukan. Dalam mengotomatisasi keputusan dengan jumlah yang banyak, *AutoGluon* memungkinkan pengembang untuk membuat model jaringan saraf yang berkinerja tinggi dengan hanya tiga baris kode. Pengembang hanya membutuhkan waktu untuk menentukan kapan ingin menyelesaikan model *deep learning* tanpa perlu melakukan pengujian terhadap pengujian data secara manual (Krishnan, 2020).

Berikut adalah contoh *code* dari *AutoGluon*:

```
from autogluon import TabularPrediction as task
predictor = task.fit("train.csv", label="class")
predictions = predictor.predict("test.csv")
```

Gambar 2. 1 Ilustrasi *AutoGluon*

AutoGluon menyediakan fit() dengan opsi tambahan bagi *user* dalam mengatur waktu *training* dalam memaksimalkan hasil akurasi prediksi.

Tabel 2. 2 Opsi *Hyperparameter AutoGluon*

Model fit	Fungsi
`hyperparameter_tune = True`	Mengoptimalkan hyperparameter pada setiap model
`auto_stack = True`	Pemilihan model strategi ensembling secara adaptif
`time_limits`	Mengatur waktu runtinme
`eval_metric`	Metrik yang digunakan dalam mengevaluasi kinerja prediktif

Dalam melakukan prediksi dengan menggunakan *AutoGluon*, terdapat hasil yang didalamnya terdapat indikator yang didapatkan melalui perintah *leaderboard().* Berikut merupakan penjelasan mengenai indikator yang dihasilkan melalui hasil prediksi dengan *AutoGluon* (AutoGluon.Ai, 2023):

Tabel 2. 3 Indikator Penjelasan *AutoGluon*

Indikator Prediksi	Keterangan
Model	Nama model yang dihasilkan
Score_val	Skor validasi model yang dihasilkan untuk *eval_metric*
Eval_metric	Metrik evaluasi yang digunakan untuk menghitung nilai validasi
Pred_time_val	Waktu yang dibutuhkan untuk menarik kesimpulan (inferensi) untuk menghitung prediksi berdasarkan data validasi dari awal hingga akhir
Fit_time	Durasi waktu yang dibutuhkan dalam melakukan melatih model dari awal hingga akhir

Indikator Prediksi	Keterangan
Pred_time_val_marginal	Informasi waktu tambahan yang digunakan untuk menarik kesimpulan (inferensi) dalam menghitung prediksi untuk nilai data validasi
Fit_time_marginal	Informasi waktu tambahan yang digunakan dalam mengetahui penyesuaian durasi waktu yang dibutuhkan dalam melatih model
Stack_level	Tumpukan tingkat model yang dihasilkan
Can_infer	Kemampuan model dalam menarik kesimpulan pada data baru
Fit_order	Urutan berdasarkan kesesuaian model

Untuk mengetahui hasil nilai akurasi pada hasil prediksi dengan *AutoGluon*, berikut merupakan rincian hasil yang didapatkan melalui perintah *evaluated():*

1. Accuracy

   Accuracy atau akurasi adalah jumlah data yang dilakukan prediki dengan benar yang dibandingkan dengan jumlah data secara keseluruhan (Yuli Sun Hariyani, 2020). Berikut merupakan rumus perhitungan dari akurasi :

   $$Akurasi = \frac{TP+TN}{TP+TN+FP+FN}$$

   Keterangan:

   TP = *True Positive*

   TN = *True Negative*

   FP = *False Positive*

   FN = *False Negative*

2. Balaced accuracy

   Balanced accuracy atau keseimbangan akurasi digunakan sebagai patokan dalam mengukur kefektifan terhadap model dengan data yang tidak seimbang (Reni Amelia, 2022). Berikut merupakan rumus perhitungan dari balaanced accuracy:

   $$balanced\ accuracy = \frac{Sensitivity+specificity}{2}$$

$$Sensitivity = \frac{A}{A+C}$$

$$Specificity = \frac{D}{D+B}$$

Keterangan:

A = kolom nilai A

B = kolom nilai B

C = kolom nilai C

D = kolom nilai D

3. MCC

   MCC (*Matthew Correlation Coefficient*) adalah metode yang digunakan dalam menilai kinerja algoritma klasifikasi berdasarkan perhitungan *confusion matrix* (Novia Hasdyna, 2020). Berikut merupakan rumus perhitungan dari MCC:

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

4. ROC-AUC

   ROC (*Receiver Operating Characteristic*) digunakan sebagai visualisasi dalam menilai akurasi dalam melakukan klasifikasi serta membandingkan dengan berbagai model klasifikasi (Vercellis, 2009). Kemudian (*Area Under the* Curve) digunakan dalam melakukan perkiraan nilai probabilitas output yang berasaal dari sampel untuk menilai kinerja diskriminatif yang dilakukan pemilahan dengan acak melalui positif maupun negatif (Hastuti, 2012)

5. F1

   F1 Score merupakan nilai yang diberikan pembobotan berupa perbandingan antara nilai presisi atau *precision* dengan *recall* (Mawadatul Maulidah, 2020)

$$F1\ score = 2x\frac{Precision\ x\ Recall}{(Precision + Recall)}$$

6. *Precision*

   Precision adalah presentase fakta maupun kesesuaian antara data yang memiliki kesesuaian yang tepat dengan hasil prediksi positif (Maximillian Christianto, 2020). Berikut merupakan rumus perhitungan dari *precision:*

$$precision = \frac{TP}{(TP + FP)}$$

7. *Recall*

   Recall merupakan metode yang digunakan untuk menilai pola positif yang dilakukan klasifikasi secara benar dalam data kelompok (Maximillian Christianto, 2020). Berikut merupakan rumus perhitungan dari *recall*:

$$recall = \frac{TP}{(TP + FN)}$$

Berikut adalah berbagai tugas yang dapat dikerjakan oleh *developer machine learning* yaitu:

1. *Tabular Prediction*

   AutoGluon dapat membuat model dalam memprediksi nilai untuk satu kolom berdasarkan nilai kolom lain dalam kumpulan data standar umum yang di representasikan dalam bentuk tabel (biasanya disimpan dalam bentuk CSV).

2. *Image Prediction*

   AutoGluon melakukan model klasifikasi gambar dengan kualitas tinggi berdasarkan kontennya.

3. *Object Detection*

   AutoGluon akan menentukan kebradaan dan posisi objek dalam gambar yang dilakukan secara otomatis untuk menghasilkan pola deteksi objek.

4. *Text Prediction*

   Prediksi teks yang dihasilkan oleh *AutoGluon* secara otomatis

5. *Multimodal Prediction*

   Prediksi dengan model *multimodal prediction* dilakukan dalam waktu yang bersamaan dengan mencampur jenis data numerik dan kategorikal. Prediksi ini dapat berjalan karena pada dasarnya pada model ini dilakukan pemisahan antara teks, kategori, dan angka yang nantinya akan digabungkan diantara metode tersebut.

2.1.6 Statistika Deskriptif

Statistika Deskriptif merupakan statistik yang digunakan untuk melakukan menganalisis dan mengatur data yang dapat memberikan ilustrasi mengenai keadaan suatu peristiwa secara ringkas sehingga dari peristiwa tersebut dapat diperoleh arti dari peristiwa tertentu (Kualitatif, 2016).

2.1.6.1 Rata-Rata

Mean atau dikenal dengan rata-rata nilai dalam suatu data, dimana dalam mencari nilai rata-rata dapat dilakukan dengan menentukan jumlah data yang akan dibagi dengan banyaknya jumlah data secara keseluruhan. Nilai rata-rata berasal dari berbagai jenis data yang dapat ditelusuri dengan melalui data tunggal maupun data kelompok. Berikut merupakan rumus perhitungan dari rata-rata (Mardhiyatirrahmah, 2023):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Keterangan:

Xi = nilai data

N = banyak data

2.1.6.2 Standar Deviasi

Standar deviasi atau disebut dengan simpangan baku merupakan teknik yang digunakan dalam menentukan persebaran data, jarak titik data individu, rata-rata dalam nilai data

sampel. Standar deviasi digunakan untuk memahami sampe data yang digunakan mampu menggambarkan populasi data keseluruhan. Berikut adalah rumus perhitungan dari standar deviasi (Devilia Rahmawati, 2020)

$$s = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - \left(n\sum_{i=1}^{n} x_i^2\right)^2}{n(n-1)}}$$

Keterangan:

S = nilai hasil standar deviasi atau simpangan baku

$\bar{x}$ = nilai rata-rata

xi = nilai x ke-i

n = *sample size* atau ukuran sampel

2.1.6.3 Median

Median merupakan nilai tengah dalam suatu data, dimana nilai dibagi menjadi dua bagian yang sama dalam data dengan jumlah yang sama. Kemudian nilai dalam tersebut diurutkan mulai dari nilai terkecil hingga nilai terbesar. Berikut merupakan rumus perhitungan dari median (Mardhiyatirrahmah, 2023)

$$M_e = t_b + \left(\frac{\frac{1}{2}n - f_k}{f}\right)c$$

Keterangan:

M_e = nilai hasil median

Tb = nilai tepi bawah median

n = jumlah banyak data

Fk = nilai frekuensi kumulatif

F = nilai frekuensi kelas median

C = nilai panjang kelas

2.1.6.4 Kuartil

Kuartil merupakan nilai yang digunakan untuk mengatur nilai pembagian atau *cutoff* dalam mendistribusikan nilai frekuensi. Nilai kuartil bertujuan untuk memutuskan nilai batas distribusi menjadi empat bagian dengan nilai masing-masing sebesar 25% (Moch Haris, 2023). Nilai kuartil yang umumnya digunakan yaitu nilai kuartil 1 sebesar 25%, kuartil 2 atau median sebesar 50%, dan nilai kuartil 3 sebesar 75%.

$$K_i = \frac{i(n+1)}{4}$$

Keterangan:

i = nilai data ke-i

n = jumlah banyak data

2.1.7 *Standard Scaler*

Dalam melakukan normalisasi data yang dilakukan dalam penelitian ini menggunakan teknik yaitu *Standard Scaler*. Normalisasi data yang dilakukan dengan *standard scaler* bertujuan untuk mengganti nilai yang terdapat dalam data mentah menjadi data dalam dengan bentuk nilai *mean* 0 dan nilai standar deviasi 1. Proses penggantiaan nilai data mentah membantu dalam mengantisipasi nilai variabel yang mendominasi dengan perbedaan skala yang besar (Yosiko Aditya Pratama, 2023). Keuntungan penggunaan teknik *standard scaler* yaitu dapat mengubah nilai dengan nilai atribut positif dan negatif menjadi nilai dengan distribusi yang mirip (Lucas B.V. de Amorima, 2022). Berikut merupan rumus dari *standard scaler*:

$$Z = \frac{X - \mu}{\sigma}$$

Keterangan:

Z = hasil *standard scaler*

X = nilai asil data

μ = nilai *mean* data

σ = nilai standar deviasi data

2.1.8 Confusion Matrix

Confusion matrix merupakan sebuah metode yang digunakan dalam menguji akurasi hasil klasifikasi untuk mengidentifikasi catatan berdasarkan hasil berbagai kelas dalam mendapatkan nilai akurasi (Sarbaini, 2021)**.** *Confusion matrix* digunakan dalam menunjukkan hasil evaluasi model dalam dataset dengan kelas pertama dimisalkan sebagai kelas positif dan kelas kedua dimisalkan sebagai kelas negatif (Detrinal Putra, 2020)**.** *Confusion Matrix* disajikan dalam bentuk tabel yang didalamnya berupa jumlah *data testing* dengan benar maupun salah yang telah diklasifikasikan. Berikut ini merupakan contoh dari tabel *confusion matrix* (Indriani, 2014):

Tabel 2. 4 Ilustrasi *confusion matrix*

Kelas sebelumnya	**Kelas Prediksi**		
		1	0
	1	TP	FN
	0	FP	TN

Berikut merupakan penjelasan pada tabel 2.4 diatas:

True Positive (TP) adalah total dokumen yang berasal dari kelas 1 yang dinyatakan benar serta masuk kedalam klasifikasi sebagai kelas 1

False Positive (FP) adalah total dokumen yang berasal dari kelas 0 yang dinyatakan benar serta masuk kedalam klasifikasi sebagai kelas 1

True Negative (TN) adalah total dokumen yang berasal dari kelas 0 yang dinyatakan benar serta masuk kedalam klasifikasi sebagai kelas 0

False Negative (FN) adalah total dokumen yang berasal dari kelas 1 yang dinyatakan benar serta masuk kedalam klasifikasi sebagai kelas 0

2.1.9 *ROC Curve*

ROC Curve (*Reicever Operating Characteristic*) memberikan gambaran mengenai visualisasi dalam mengevaluasi akurasi pengklasifikasian dengan membandingkan model klasifikasi yang berbeda. *ROC Curve* merupakan representasi pada plot grafis dua dimensi yang menyajikan proporsi *false positive* (fp) atau positif palsu pada sumbu horizontal dan proporsi *true positive* (tp) atau positif sebenarnya pada sumbu vertikal. Pada visualisasi grafis terdapat titik (0,1) memperlihatkan pengklasifikasian secara ideal yang tidak membuat kesalahan dalam membuat prediksi yang terjadi karena proporsi positif palsu memiliki nilai nol (fp=0) dan nilai proporsi positif sebenarnya mencapai nilai maksimum (tp=1). Pada titik (0,0) memberikan gambaran mengenai klasifikasi dalam melakukan prediksi kelas {-1} untuk seluruh observasi, kemudian titik (1,1) memiliki hubungan dengan klasifikasi dalam melakukan prediksi kelas {1} untuk seluruh observasi (Vercellis, 2009). *ROC Curve* digunakan dalam membandingkan berbagai metode *classifier* maupun model *classifier* yang memiliki parameter berbeda dalam memperoleh model yang paling ideal (Laila Qadrini, 2021). Berikut merupakan rumus dari ROC Curve (Wen Zhu, 2010):

$$TPR = \frac{TP}{(TP + FN)}$$

$$FPR = \frac{FP}{(FP + TN)}$$

Keterangan:

TPR = *True Positive Rate*

FPR = *False Positive Rate*

TP = *True Positive*

TN = *True Negative*

FP = *False Positive*

FN = *False Negative*

2.2 Kajian Literatur

Kajian literatur merupakan kajian yang berasal dari berbagai penelitian terdahulu sebagai bahan referensi bagi penulis selama penelitian.

2.2.1 Komparasi Metode *Deep Learning, Naive Bayes,* dan *Random Forest* Untuk Prediksi Penyakit Jantung

Penelitian yang dilakukan oleh Ivana Alhabib, Ahmad Faqih, Fatihanursari Dikananda yang dipublikasikan pada tahun 2022 bertujuan dalam membantu dokter dalam melakukan prediksi dalam mendiagnosis dengan tepat dan akurat supaya penanganan penyakit jantung koroner dapat dilakukan lebih awal. Algoritma yang digunakan dalam penelitian ini yaitu *Naïve Bayes* dan *Random Forest* yang memanfaatkan data rekam medis pasien dengan tujuan dalam menghitung probabilitas terkait kemungkinan seorang pasien. Penelitian ini menggunakan dataset yang diunggah di Kaggle.com dengan judul *Heart Attack Analysis & Prediction Dataset* yang dapat diketahui bahwa algoritma *deep learning* menjadi algoritma terbaik dengan nilai sebesar 83,49% dan dengan nilai AUC (*Area Under Curve*) sebesar 0,902.

2.2.2 Prediksi Pasien Dengan Penyakit Kardiovaskular Menggunakan *Random Forest*

Penelitian yang dilakukan oleh Mochammad Anshori, Nindynar Rikatsih, M. Syauqi Haris yamg dipublikasikan pada tahun 2022 bertujuan untuk membuktikan algoritma *random forest* dapat memprediksi dengan baik dibandingkan dengan metode *machine learning* lainnua. Data yang digunakan dalam penelitian ini yaitu data sekunder yang diakses melalui mendeley. Metode yang digunakan dalam penelitian ini yaitu metode *random forest* yang dikombinasikan dengan *library autoweka.* Hasil yang didapatkan pada penelitian ini yaitu algoritma *random forest* dapat diterapkan dalam pembutatan model prediksi penuakit kardiovaskular dengan nilai TPR sebesar 0,983, nilai FPR sebesar 0,024, nilai AUC sebesar 0,99969 dan akurasi sebesar 98%.

2.2.3 Perbandingan Performansi Algoritma Pengklasifikasifikasian Terpadu Untuk Kasus Penyakit Kardiovaskular

Penelitian yang dilakukan oleh Adi Nugroho, Agustinus Bimo Gumelar, Adri Gabriel Sooai, Dyana Sarvasti, Paul L Tahalele yang dipublikasikan pada tahun 2020 bertujuan untuk membandingkan kinerja dari algoritma klasifikasi yang nantinya akan dilakukan klasifikasi melalui faktor risiko penyakit kardiovaskular. Metode yang digunakan dalam penelitian ini yaitu *k-nearest neighbors, stohtastic gradient descent, random forest, neural network,* dan *logistic regression* dengan *tools* bantuan berupa *orange* dan *weka*. Data yang digunakan dalam penelitian ini yaitu data yang berasal dari *github.com* yang mendapatkan hasil berupa metode *Neural Network* menjadi metode yang terbaik dengan nilai akurasi sebesar 89,60%, nilai AUC sebesar 0,873, nilai presisi sebesar 0,877, dan nilai recall sebesar 0,896

2.2.4 Prediksi Penyakit Jantung *Cardiovascular* Menggunakan Algoritma Klasifikasi

Penelitian yang dilakukan oleh Wahyu Nugraha yang dipublikasikan pada tahun 2021 bertujuan untuk menciptakan model prediksi dengan algoritna klasifikasi pada *machine learning* untuk memprediksi penyakit kardiovaskular untuk mengetahui model prediksi terbaik dan akurat. Data yang digunakan dalam penelitian ini yaitu berasal dari *Kaggle.com idan* biorxiv.*org* dengan metode algoritma *random forest, support vector machines, gradient boosting machines, xgboost, light gbm*. Hasil yang didapatkan pada penelitian ini yaitu model klasifikasi dengan algoritma *xgboost* mendapatkan nilai yang tertinggi jika dibandingkan algoritma yang lainnya dengan nilai *Accuracy* sebesar 80%, nilai *F1-Score* sebesar 86%, dan nilai *AUC* sebesar 75%.

2.2.5 Komparasi *Support Vector Machine, Logistic Regression,* Dan *Artificial Neural Network* dalam Prediksi Penyakit Jantung

Penelitian yang dilakukan oleh Fitri Handayani, Kartika Sari Kusuma, Hedy Leoni Asbudi, Rona Guines Purnasiwi, Reti Kusuma, Andi Sunyoto, Windha Mega Pradnya yang dipublikasikan pada tahun 2021 yang bertujuan untuk mencapai tingkat akurasi yang terbaik dengan ketiga metode yang digunakan dalam berbagai data. Data yang digunakan dalam penelitian ini yaitu data yang berasal dari *Uci machine learning* yang

dikombinasikan dengan menggunakan metode *Support Vector Machine, Logistic Regression,* Dan *Artificial Neural Network*. Hasil yang didapatkan dalam penelitian ini yaitu dengan komposisi pada data latih dah data uji yang berbeda maka metode algoritma yang terbaik juga akan berbeda. Apabila dengan komposisi 90:10 maka metode yang terbaik adalah *Logistic Regression,* Dan *Artificial Neural Network* dengan nilai sebesar 80% . Jika menggunakan komposisi 80:20 maka metode yang terbaiik adalah *Logistic Regression* dengan nilai sebesar 80%. Jika menggunakan komposisi 70:30 maka metode terbaik adalah *Artificial Neural Network* dengan nilai sebesar 82%. Apabila menggunakan kompoisisi 60:40 maka metode yang terbaik adalah *Logistic Regression,* dengan nilai akurasi sebesar 83%.

2.2.6 *A Data-driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning*

Penelitian yang dilakukan oleh An Dinh, Stacey Miertschin, Amber Young, Somya D.Mohanty yang dipublikasikan pada tahun 2019 bertujuan untuk berupaya dalam mengemabangkan model yang berbasis data untuk mengetahui karakteristik dasar. Data yang digunakan dalam penelitian ini yaitu menggunakan data *National Health and Nutrition Examination Survey* yang diolah menggunakan metode *logistic regression, support vector machine, random forest, xgbosst, ensemble*. Hasil yang didapatkan pada penelitian ini yaitu pada model *ensemble* dalam penyakit kardiovaskular mendapatkan hasil AU-ROC sebesar 83,1% tanpa hasil laboratorium dan 83,9% dengan hasil laboratorium. Dalam penyakit diabetes dengan model *xgboost* didapatkan hasil AU-ROCsebesar 86,2% tanpa laboratorium dan 95,7% jika dengan laboratorium.

2.2.7 *Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques*

Penelitian yang dilakukan oleh Rajkumar Gangappa Nadakinamani, A.Reyana, Sandeep Kaustish, A.S. Vibith, Yogita Gupta, Sayed F. Abdelwahab, Ali Wagdy Mohamed yang dipublikasikan pada tahun 2022 bertujuan untuk melakukan prediksi penyakit kardiovaskular untuk mendapatkan hasil yang lebih akurat dengan waktu yang lebih singkat. Data yang digunakan dalam penelitian ini yaitu dataset yang berasal dari *National*

Health and Nutrition yang diolah menggunakan algoritma model *REP Tree, M5P Tree, Random Tree, Linear Regression, Naïve Bayes,* J48, dan *JRIP* untuk mengklasifikasikan data. Hasil yang didapatkan dalam penelitian ini yaitu pada model *Random Tree* dengan nilai *accuracy* sebesar 100%, nilai MAE terendah sebesah 0,0011, nilai RMSE terendah sebesar 0,0231, serta waktu prediksi tercepat 0,01 detik.

2.2.8 *Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis and Comparison*

Penelitian yang dilakukan oleh Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian M.W. Quinn, Mohammad Ali Moni yang dipublikasikan pada tahun 2021 yang bertujuan untuk menemukan teknik *machine learning* yang terbaik berdasarkan algoritma yang diterima. Data yang digunakan dalam penelitian ini yaitu dataset yang diakses melalui *Kaggle.com* yang diolah menggunakan algoritma *linear regression, adaboostm1, multilayer perceptron, k-nearest neighbor, decision tree,* dan *random forest*. Hasil penelitian yang didapatkan yaitu algoritma klasifikasi yang dapat bekerja dengan baik terdapat pada algoritma *k-nearest neighbor, decision tree,* dan *random forest* yang masing-masing mendapatkan nilai *accuracy* sebesar 100%.

2.2.9 *A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques*

Penelitian yang dilakukan oleh Raja Krishnamoorthi, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C,Kalpana, Basant Tiwari yang dipublikasi tahun 2022 yang bertujuan untuk mengetahui bagaimana analitik data dan teknik yang digunakan dalam *machine learning*. Data yang digunakan dalam penelitian ini yaitu dengan menggunakan dataset *Pima Indian Diabetes Database* yang diolah menggunakan metode algoritma *logistic regression, k-nearest neighbor, random forest,* dan *support vector machine*. Hasil penelitian yang diperoleh pada penelitian ini yaitu metode terbaik dalam melakukan prediksi berupa metode *logistic regression* dengan nilai 86%.

2.2.10 *Prediction of Cardiovascular Diseases based on Machine Learning*

Penelitian yang dilakukan oleh Weicheng Sun, Ping Zhang, Zilin Wang, Dongxu Li yang dipublikasikan pada tahun 2021 yang bertujuan untuk mengklasifikasikan non-linier dalam melakukan prediksi penyakit kardiovaskular sebagai referensi untuk mencegah dan mengobati penyakit kardiovaskular. Data yang digunakan dalam penelitian ini yaitu data percobaan penelitian yang berasal dari *Svetlana ulianova* yang diolah menggunakan *support vector machine, logical regression,* dan *random forest*. Hasil yang didapatkan dalam penelitian ini yaitu metode *support vector machine* menjadi metode yang terbaik dibandingkan dengan metode *random forest* dan *logica regression* dengan nilai sebesar 78,84%.

2.2.11 *FinLex : An Effective Use of Word Embeddings for Financial Lexicon Generation*

Penelitian yang dilakukan oleh Sanjiv R.Das, Michele Donini, Muhammad Bilal Zafar, John He, Krishnaram Kenthapadi yang dipublikasikan pada tahun 2021 yang bertujuan untuk menyusun dan mempertimbangkan dalam mengklasifikasikan *machine learning* yang berbeda-beda pada kumpulan data. Data yang digunakan dalam penelitian ini yaitu menggunakan dataset yang berasal dari *Kaggle.com* dan *Financial Phrase Bank* (FPB) yang diolah dengan metode *autogluon*. Hasil yang didapatkan dalam penelitian ini yaitu hasil klasifikasi dengan *Loughran-McDonald* bekerja dengan baik dibandingkan dengan klasifikasi *inverse document frequency* dengan pendekatan yang digunakan.

2.2.12 *AutoGluon : A Revolutionary Framework for Landslide Hazard Analysis*

Penelitian yang dilakukan oleh WenWen Qi, Chong Xu, Xiwei Xu yang dipublikasikan pada tahun 2021 yang bertujuan untuk mencapai hasil klasifikasi atau regresi pada setiap kolom data. Data yang digunakan dalam penelitian ini yaitu dataset yang berasal dari dataset yang digunakan dalam penelitian (Yingying Tian, 2018) yang diolah dengan menggunakan metode *AutoGluon*, khususnya pada *AutoGluon Tabular*.Hasil yang didapatkan dalam penelitian ini yaitu hasil terbaik dalam model *machine learning* yang diukur pada nillai ROC-AUC adalah sebesar 0,94 yang membutuhkan waktu pemrosesan selama 47,33 detik.

2.2.13 *AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data*

Penelitian yang dilakukan oleh Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, Alexander Smola yang dipublikasikan pada tahun 2020 yang bertujuan dengan fokus untuk menggabungkan beberapa model dan menyusun menjadi beberapa lapisan melalui pemilihan model/hypermater. Data yang digunakan dalam penelitian ini yaitu dengan dataset pada *OpenML AitoML* dan *Kaggle* yang diolah dengan menggunakan metode *AutoGluon,* khususnya *AutoGluon-Tabular*. Hasil yang didapatkan dalam penelitian ini yaitu mendapatkan hasil yang lebih akurat dibandingkan dengan *framework AutoML* yang lainnya dengan nilai sebesar Avg. Rank sebesar 1,9234 dan Avg. Rescaled Loss sebesar 0,1660.

2.2.14 *AutoGluon-TimeSeries*: *AutoML for Probabilistic Time Series Forecasting*

Penelitian yang dilakukan oleh Oleksandr Schur, Caner Turkmen, Nick Erickson, Huibin Shen, Alexander Shirkov, Tony Hu, Yuyang Wang yang dipublikasikan pada tahun 2023 yang bertujuan dalam mengevaluasi hasil akurasi estimasi probabilistik *AutoGluon-Time Series* dengan menggunakan metode *forecasting* yang berbasis *machine learning*. Metode yang digunakan dalam penelitian ini yaitu *AutoGluon* dengan hasil yang didapatkan yaitu kinerja kuat yang ditunjukkan pada *AutoGluon-TimeSeries* yang dapat mengungguli pada metode *forescasting* mengenai akurasi titik dan kuantisasi, bahkan sering mengungguli dalam kombinasi yang terbaik dibandingkan metode sebelumnya.

2.2.15 *Forecasting the Walking Assistance Rehabilitation Level of Stroke Patients Using Artificial Intelligence*

Penelitian yang dilakukan oleh Kanghyeon Seo, Bokjin Chung, Hamsa Priya Panchaseelan, Taewoo Kim, Hyejung Park, Byungmo Oh, Minho Chun, Sunjae Won, Donkyu Kim, Jaewon Beom, Doyounh Jeon, Jihoon Yang yang dipublikasikan pada tahun 2021 yang bertujuan untuk mengevaluasi efektivitas dalam penggunanaaan berbagai jenis model klasifikasi *machine learning* dan *deep learning*. Data yang digunakan dalam penelitian ini yaitu dataset yang berasal dari 5 rumah sakit *Chung-Ang University Hospital*, *Seoul National University Hospital, National Traffic Injury Rehabilitation Hospital*, *The Catholic University of Korea Yeouido St. Marry's Hospital*, dan *Asan Medical Center* yang dilakukan dengan menggunakan metode *AutoGluon*.

Sehingga dalam peneliitan ini menghasilkan nilai klasifikasi kinerja prediksi tertinggi hamper 92% pada *recall, precision,* dan *F1-Score* serta 86,8% pada *accuracy.*

Tabel 2. 5 Tabel Kajian Induktif

No	Penulis, Tahun	Objek Penelitian	Metode									
			AutoGluon	Naïve Bayes	Random Forest	Logistic Regression	K-NN	SVM	NN	Light Gbm	XGBoost	ANN	DT
1	(Ivana Alhabib, 2022)	Penaganan Penyakit Jantung Koroner		✓	✓								
2	(Mochammad Anshori, 2022)	Faktor risiko penyakit kardiovaskular			✓	✓							
3	(Adi Nugroho, 2020)	Klasifikasi penyakit kardiovaskular			✓	✓	✓	✓					
4	(Nugraha, 2021)	Klasifikasi penyakit kardiovaskular			✓		✓			✓	✓		

No	Penulis, Tahun	Objek Penelitian	Metode									
			AutoGluon	Naïve Bayes	Random Forest	Logistic Regression	K-NN	SVM	NN	Light Gbm	XGBoost	ANN	DT
5	(Fitri Handayani, 2021)	Prediksi penyakit jantung				✓		✓				✓	
6	(An Dinh, 2019)	Prediksi penyakit diabetes dan kardiovaskular			✓	✓		✓			✓		
7	(Rajkumar Gangappa Nadakinamani, 2022)	Pengembangan model prediksi			✓	✓		✓			✓		
8	(Md Mamun Ali, 2021)	Prediksi Penyakit Jantung			✓	✓	✓						✓

No	Penulis, Tahun	Objek Penelitian	Metode									
			AutoGluon	Naïve Bayes	Random Forest	Logistic Regression	K-NN	SVM	NN	Light Gbm	XGBoost	ANN	DT
9	(Raja Krishnamoorthi, 2022)	Prediksi penyakit diabetes			✓	✓	✓	✓					
10	(Weicheng Sun, 2021)	Prediksi penyakit kardiovaskular			✓	✓		✓					
11	(Sanjiv R. Da, 2022)	Klasifikasi machine learning pada Kumpulan data	✓										
12	(Wenwen Qi, 2021)	Klasifikasi dan regersi setiap kolom data	✓										
13	(Nick Erickson, 2020)	Penggabungan beberapa	✓										

No	Penulis, Tahun	Objek Penelitian	Metode									
			AutoGluon	**Naïve Bayes**	**Random Forest**	**Logistic Regression**	**K-NN**	**SVM**	**NN**	**Light Gbm**	**XGBoost**	**ANN**	**DT**
		model menjadi beberapa lapisan											
14	(Oleksandr Shchur, 2023)	Evaluasi Hasil Estimasi Probabiilistik	✓										
15	(Kanghyeon Seo, 2021)	Evaluasi efektivitas model klasifikasi	✓										

Posisi penelitian : *AutoGluon, Random Forest, K-NN*

BAB III

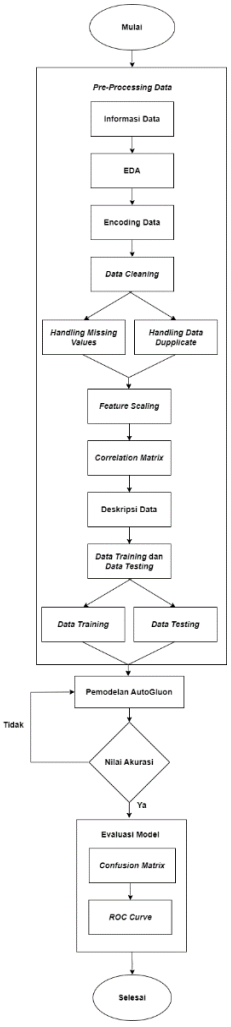
METODE PENELITIAN

3.1 Objek Penelitian

Objek pada penelitian kali ini yaitu dataset yang diakses melalui situs *Kaggle.com* dengan implementasi *AutoGluon* dalam dataset *Cardiovascular Risk Disease Prediction Dataset* dengan fokus pada penyakit jantung.

3.2 Diagram Alur Penelitian

Berikut merupakan diagram alur penelitian yang menunjukkan alur tahapan dalam penelitian dari awal hingga akhir penelitian sebagai berikut:



Gambar 3. 1 Diagram Alur Penelitian

3.3 Pre-processing Data

Pre-processing data dilakukan untuk mengubah data menjadi format yang sederhana dan efisien dalam memenuhi kebutuhan pengguna (Saifullah, 2017). *Pre-processing data* yang dilakukan pada tahap ini yaitu untuk menyiapkan data agar siap untuk digunakan dalam menjalankan pemodelan data. Berikut merupakan langkah *pre-processing data* yang dilakukan dalam penelitian ini:

3.3.1 Informasi Data

Data info merupakan tahapan yang dilakukan dalam penelitian ini untuk mengetahui informasi yang terdapat dalam dataset. Informasi yang terdapat dalam dataset digunakan untuk membantu mengumpulkan informasi yang dibutuhkan dalam penelitian. Variabel data yang digunakan dalam penelitian ini yaitu *General_Health, Checkup, Exercise, Heart_Disease, Skin_Cancer, Other_Cancer, Depression, Diabetes, Arthritis, Sex, Age_Category, Height_(cm), Weight_(kg), BMI, Smoking_History, Alcohol_Consumption, Fruit_Consumption, Green_Vegetables_Consumptions,* dan *FriedPotato_ Consumptions*. Variabel data yang digunakan sebagai target dalam melakukan prediksi yaitu variabel *Heart_Disease*, pemilihan variabel *Heart_Disease* sebagai target sesuai dengan tujuan penelitian yaitu untuk melakukan prediksi terhadap penyakit jantung.

3.3.2 *Exploratory Data Analysis* (EDA)

Exploratory Data Analysisis (EDA) digunakan untuk mengeksplorasi berbagai pola data yang disajikan dalam dataset. Eksplorasi dilakukan dengan melakukan visualisasi data dengan jenis *bar chart* pada setiap variabel.

3.3.3 *Encoding Data*

Encoding Data dilakukan dengan mengubah nilai yang terdapat dalam variabel pada *dataset* dari nilai kategorikal menjadi nilai numerik. Penggunaan nilai numerik dilakukan untuk memudahkan dalam menjalankan pemodelan data.

3.3.4 *Data Cleaning*

Data cleaning dilakukan untuk membersihkan data agar data siap digunakan dalam menjalakan pemodelan data. Langkah *data cleaning* yang dilakukan pada tahap ini yaitu dengan melakukan *handling missing values, handling duplicate*, dan *feature scaling*.

3.3.5 Matriks Korelasi

Matriks korelasi atau *correlation matrix* dilakukan untuk mengetahui hubungan antara satu variabel dengan variabel yang lainnya. Tahap ini untuk mencari variabel yang memiliki hubungan yang kuat dengan variabel yang lain.

3.3.6 Deskripsi Data

Deskripsi data atau *Data describe* dalam penelitian ini dilakukan untuk mengetahui isi gambaran secara ringkas pada dataset. Ringkasan dataset yang disajikan berupa nilai statistik dasar dengan nilai numerik.

3.3.7 *Data Training dan Data Testing*

Data training data dan *data testing* pada tahap ini dilakukan untuk melakukan pelatihan pada *dataset* dan juga untuk menguji *dataset* dengan menggunakan *dataset* yang sebelumnya telah dilakukan *split data* atau pemisahan data. Hal ini bertujuan agar data yang digunakan telah siap untuk melakukan pemodelan maupun melakukan evaluasi hasil model data.

3.4 Pemodelan *AutoGluon*

Pemodelan *AutoGluon* merupakan *open-source* yang dikembangkan oleh *Amazon Web Series* dalam menyederhanakan proses pemodelan dan pemilihan model terkait *machine learning*. *AutoGluon-Tabular* merupakan model dalam *machine learning* yang akurat dalam dataset tabular yang belum diproses dengan mampu meraktit beberapa model dan menyusun menjadi beberapa lapisan (Amazon Sage Maker, 2023). Pemodelan AutoGluion yang dilakukan pada tahap ini untuk melakukan klasifikasi data untuk mencari nilai akurasi. Dalam melakukan pemodelan dengan *AutoGluon* apabila nilai akurasi yang dihasilkan tidak sesuai harapan atau mengalami kendala, maka melakukan penyesuaian ulang *hyperparameter* pada pemodelan *AutoGluon*. Apabila nilai akurasi memenuhi harapan dan tidak mengalami kendala dapat dilanjutkan dalam melakukan evaluasi model.

3.5 Evaluasi Model

Model Evaluation atau evaluasi model dilakukan dalam menganalisis hasil performa berdasarkan algoritma *machine learning* (Agus Ambarwari, 2020). Evaluasi model yang dilakukan dalam penelitian ini dengan menggunakan *confusion matrix* dan *ROC Curve*.

BAB IV

PENGUMPULAN DAN PENGOLAHAN DATA

4.1 Pengumpulan Data

Pengumpulan data untuk penelitian ini yaitu dengan data sekunder yang diakses melalui situs *Kaggle.com* berupa data hasil survey yang dilakukan oleh *Centers for Disease Control and Prevention* (CDC) dalam kegiatan *Behavioral Risk Factor Surveillance System* (BRFSS) pada tahun 2021 dengan judul dataset "*Cardiovascular Diseases Risk Prediction Dataset"* yang diakses pada tanggal 20 Juli 2023 melalui alamat situs *link* sebagai berikut : (https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset). Data tersebut kemudian diunduh dalam bentuk zip yang selanjutnya akan diekstrak kedalam format CSV dengan 19 variabel. Berikut variabel dan deskripsi data pada dataset berikut:

Tabel 4. 1 Variabel Data

Variabel	Deskripsi
General_Health	Kondisi Kesehatan responden secara umum
Checkup	lama waktu mengunjungi dokter untuk pemeriksaan rutin
Exercise	Aktivitas fisik yang dilakukan responden selama sebulan rutin selain pekerjaan rutin
Heart_Disease	Responden yang mengalami penyakit jantung koroner atau bukan

Variabel	Deskripsi
Skin_Cancer	Responden yang mengalami kanker kulit
Other_Cancer	Responden yang mengalami jenis kanker yang lain
Depression	Responden yang menderita depresi
Diabetes	Responden yang menderita diabetes
Arthritis	Responden yang mengalami arthritis
Sex	Jenis kelamin responden
Age_Category	Kategori umur responden
Height_(cm)	Tinggi badan responden
Weight_(kg)	Berat badan responden
BMI	Masa tubuh responden
Smoking_History	Pengalaman merokok responden
Alcohol_Consumption	Pengalaman konsumsi alkohol responden
Fruit_Consumption	Pengalaman konsumsi buah responden
Green_Vegetables_ Consumptions	Pengalaman konsumsi sayuran hijau responden
FriedPotato_ Consumption	Prengalaman konsumsi kentang goreng responden

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_(cm)	Weight_(kg)	BMI	Smoking_History	A
0	Poor	Within the past 2 years	No	No	No	No	No	No	Yes	Female	70-74	150.0	32.66	14.54	Yes	
1	Very Good	Within the past year	No	Yes	No	No	No	Yes	No	Female	70-74	165.0	77.11	28.29	No	
2	Very Good	Within the past year	Yes	No	No	No	No	Yes	No	Female	60-64	163.0	88.45	33.47	No	
3	Poor	Within the past year	Yes	Yes	No	No	No	Yes	No	Male	75-79	180.0	93.44	28.73	No	
4	Good	Within the past year	No	No	No	No	No	No	No	Male	80+	191.0	88.45	24.37	Yes	

Gambar 4. 1 Tampilan 5 data teratas

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_(cm)	Weight_(kg)	BMI	Smoking_Hist
308849	Very Good	Within the past year	Yes	No	No	No	No	No	No	Male	25-29	168.0	81.65	29.05	
308850	Fair	Within the past 5 years	Yes	No	No	No	No	Yes	No	Male	65-69	180.0	69.85	21.48	
308851	Very Good	5 or more years ago	Yes	No	No	No	Yes	Yes, but female told only during pregnancy	No	Female	30-34	157.0	61.23	24.69	
308852	Very Good	Within the past year	Yes	No	No	No	No	No	No	Male	65-69	183.0	79.38	23.73	
308853	Excellent	Within the past year	Yes	No	No	No	No	No	No	Female	45-49	160.0	81.19	31.71	

Gambar 4. 2 Tampilan 5 data terbawah

4.2 Pre-processing Data

Sebelum dilakukan pemrosesan pada data, data perlu dilakukan penataan agar data dapat lebih mudah untuk dilakukan pengolahan data. Sehingga perlu dilakukan *preprocessing data* supaya data yang akan diolah sesuai dengan kebutuhan karena tidak semua data akan diolah (Rerung, 2018). *Pre-Processing* yang dilakukan pada penelitian ini dilakukan dengan bantuan *software Google Collaboratory* dengan Bahasa pemrograman *Python* yang kemudian dilakukan installasi autogluon dan beberapa library yang mendukung dalam *pre processing data*. Berikut merupakan *library* yang digunakan dalam penelitian ini:

1. Pandas : *library open source* yang menyajikan analisis struktur data dan manipulasi data.
2. Numpy : *library open source* yang menyajikan *array* yang efisien dan berkinerja tinggi.
3. Matplotlib : *library* yang digunakan dalam melakukan visualisasi dengan berbagai jenis grafik.

4. Seaborn : *library* yang digunakan dalam melakukan visualisasi berdasarkan hasil matplotlib.
5. AutoGluon : *library open source* yang digunakan dalam melakukan pemodelan otomatisasi pada *machine learning*.
6. Sklearn : *library open source* yang digunakan dalam melakukan pemisahan data maupun pemrosesan model.

4.2.1 Informasi Data

Informasi Data diakses melalui fungsi *Data Info* yang merupakan salah satu fungsi pada *library pandas* yang memberikan gambaran secara singkat mengenai data yang akan di analisis (Ranjan, 2021). Penulusuran informasi pada tahap ini untuk memperoleh informasi yang terdapat dalam *dataframe* seperti jumlah kolom, jenis data, dan jumlah baris. Berikut ini merupakan informasi yang terdapat dalam *dataframe*:

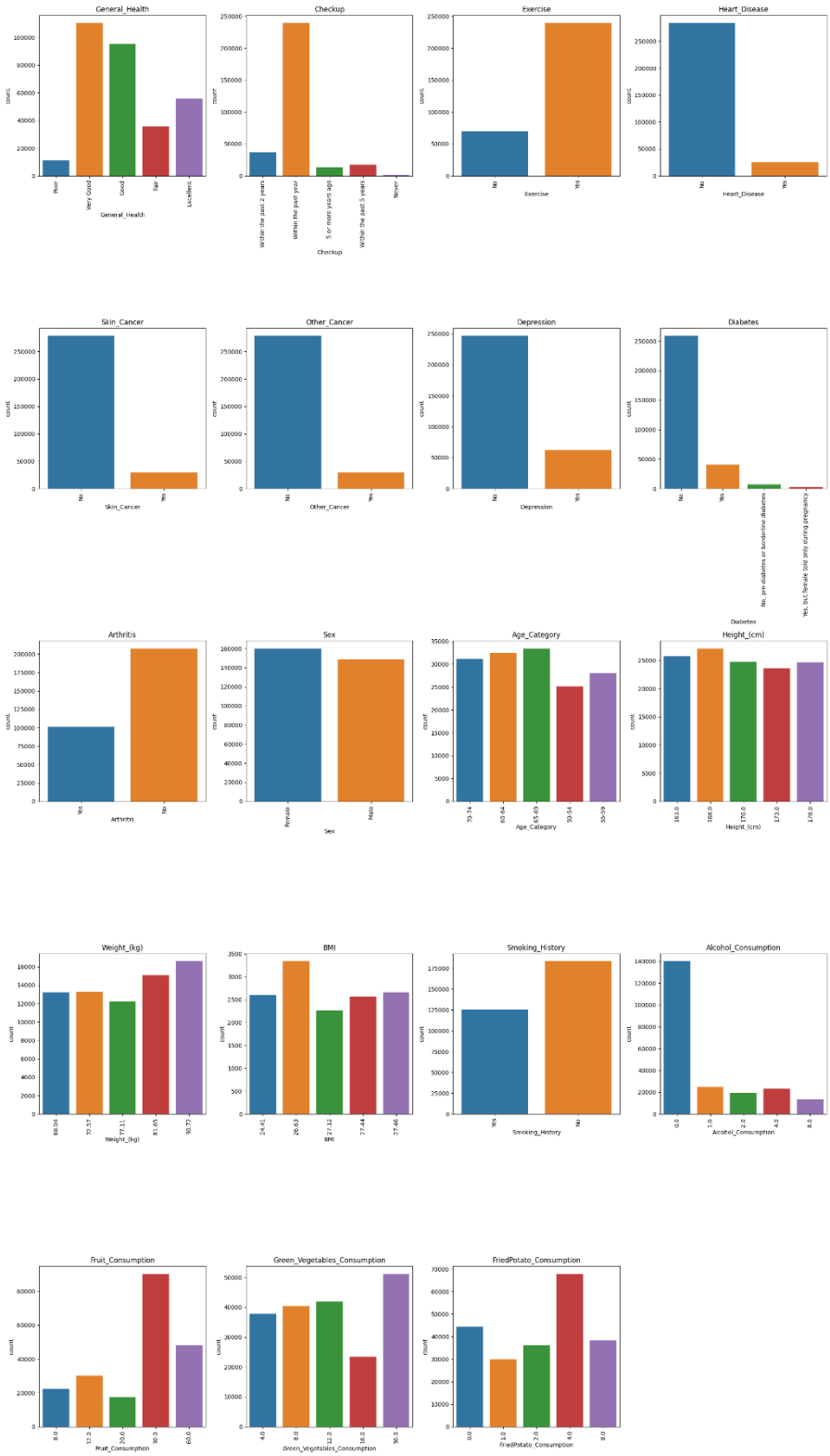
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 308854 entries, 0 to 308853
Data columns (total 19 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   General_Health                308854 non-null  object
 1   Checkup                       308854 non-null  object
 2   Exercise                      308854 non-null  object
 3   Heart_Disease                 308854 non-null  object
 4   Skin_Cancer                   308854 non-null  object
 5   Other_Cancer                  308854 non-null  object
 6   Depression                    308854 non-null  object
 7   Diabetes                      308854 non-null  object
 8   Arthritis                     308854 non-null  object
 9   Sex                           308854 non-null  object
 10  Age_Category                  308854 non-null  object
 11  Height_(cm)                   308854 non-null  float64
 12  Weight_(kg)                   308854 non-null  float64
 13  BMI                           308854 non-null  float64
 14  Smoking_History               308854 non-null  object
 15  Alcohol_Consumption           308854 non-null  float64
 16  Fruit_Consumption             308854 non-null  float64
 17  Green_Vegetables_Consumption  308854 non-null  float64
 18  FriedPotato_Consumption       308854 non-null  float64
dtypes: float64(7), object(12)
memory usage: 44.8+ MB
```

Gambar 4. 3 *Data Info*

4.2.2 *Exploratory Data Analysis* (EDA)

Exploratory Data Analysis (EDA) merupakan salah satu cara yang digunakan untuk mengeksplorasi mengenai sekumpulan data dengan menggunakan metode visualisasi grafis dalam menemukan pola data tersebut (Urminder Singh, 2020). Penggunaan

Exploratory Data Analysis bertujuan untuk menampilkan visualisasi data agar dapat dengan mudah untuk dipahami serta memberikan informasi secara keseluruhan berdasarkan Kumpulan data (Irawan, 2004). Dalam membuat visualisasi dengan menggunakan *Exploratory Data Analysis*, dilakukan dengan mendefinisikan jenis data yang digunakan. Penelitian ini menggunakan jenis data yaitu *object* dan *float64* berdasarkan informasi data yang diperoleh sebelumnya. Kemudian membuat visualisasi menggunakan *barchart* dengan 1 variabel menampilkan 5 atribut teratas apabila dalam 1 variabel terdapat 5 dari atribut. Selanjutnya menghapus *subplot* yang tidak digunakan dalam membuat visualisasi. Berikut ini merupakan hasil visualisasi dengan menggunakan *Exploratory Data Analysis* :



Gambar 4. 4 Visualisasi *Exploratory Data Analysis*

4.2.3 *Encoding Data*

Data yang terdapat dalam dataset masih dalam berbentuk kategorikal, sehingga data perlu diubah menjadi numerik dengan menggunakan teknik label encoder (Abdurraman, 2023). Penelitian ini menggunakan encoding dengan jenis label encoding yang mengubah nilai kategori menjadi nilai dengan nilai bulat (Heri Santoso, 2023). Dalam melakukan encoding data, setiap variabel perlu dilakukan pendefinisian untuk memudahkan sistem dalam membaca variabel mana yang diubah nilainya. Kemudian sistem membaca variabel yang telah didefinisikan sebelumnya dengan nilai kategori menjadi nilai numerik. Berikut ini merupakan hasil dari label encoding*:*

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_(cm)	Weight_(kg)	BMI	Smoking_History
0	3	2	0	0	0	0	0	0	1	0	10	150.0	32.66	14.54	1
1	4	4	0	1	0	0	0	2	0	0	10	165.0	77.11	28.29	0
2	4	4	1	0	0	0	0	2	0	0	8	163.0	88.45	33.47	0
3	3	4	1	1	0	0	0	2	0	1	11	180.0	93.44	28.73	0
4	2	4	0	0	0	0	0	0	0	1	12	191.0	88.45	24.37	1

Gambar 4. 5 Hasil *Encoding Data*

4.2.4 *Data Cleaning*

Data Cleaning digunakan untuk menyelesaikan permasalahan pada data yang terjadi karena kesalahan yang terjadi akibat manusia maupun mesin (Obaid Alotaibi, 2023). *Data Cleaning* yang dilakukaan dalam penelitian ini bertujuan untuk memastikan data bersih sebelum melakukan proses pemodelan *AutoGluon*. Langkah *data cleaning* yang dilakukan dalam tahap ini yaitu *handling missing values, handling duplicate,* dan *Feature Scaling*.

4.2.4.1 *Handling Missing Values*

Penanganan data yang hilang atau *missing values* dalam penelitian ini dilakukan dengan menggunakan *isnull*. *Isnull* yang digunakan dalam penelitian ini untuk memeriksa nilai yang hilang dalam data. Berikut ini merupakan jumlah data yang hilang dalam data diatas:

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_(cm)	Weight_(kg)	BMI	Smoking_History
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	..
308849	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
308850	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
308851	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
308852	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
308853	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals

308854 rows × 19 columns

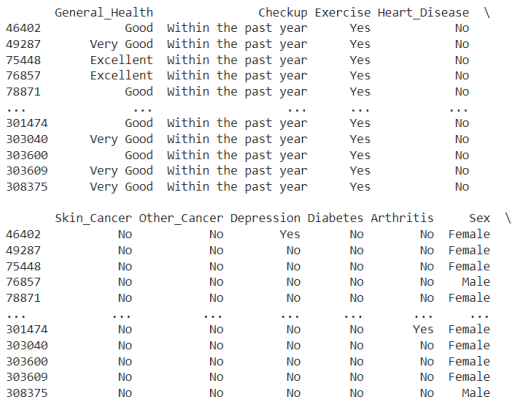
Gambar 4. 6 Memeriksa nilai data yang kosong

```
General_Health                0
Checkup                       0
Exercise                      0
Heart_Disease                 0
Skin_Cancer                   0
Other_Cancer                  0
Depression                    0
Diabetes                      0
Arthritis                     0
Sex                           0
Age_Category                  0
Height_(cm)                   0
Weight_(kg)                   0
BMI                           0
Smoking_History               0
Alcohol_Consumption           0
Fruit_Consumption             0
Green_Vegetables_Consumption  0
FriedPotato_Consumption       0
dtype: int64
```

Gambar 4. 7 Nilai data yang kosong

4.2.4.2 *Handling Dupplicate*

Penanganan data yang mengalami duplikasi atau *duplicated* dilakukan untuk memeriksa baris yang tersalin atau tergandakan. Berikut ini merupakan jumlah data yang mengalami terduplikasi diatas:



	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex
46402	Good	Within the past year	Yes	No	No	No	Yes	No	No	Female
49287	Very Good	Within the past year	Yes	No	No	No	No	No	No	Female
75448	Excellent	Within the past year	Yes	No	No	No	No	No	No	Female
76857	Excellent	Within the past year	Yes	No	No	No	No	No	No	Male
78871	Good	Within the past year	Yes	No	No	No	No	No	No	Female
...	...	...	...	...	...	...	...	...	...	...
301474	Good	Within the past year	Yes	No	No	No	No	No	Yes	Female
303040	Very Good	Within the past year	Yes	No	No	No	No	No	No	Female
303600	Good	Within the past year	Yes	No	No	No	No	No	No	Female
303609	Very Good	Within the past year	Yes	No	No	No	No	No	No	Female
308375	Very Good	Within the past year	Yes	No	No	No	No	No	No	Male

Gambar 4. 8 Data Yang Terduplikasi

```
Jumlah data yang terduplikat sebanyak: 80
```

Gambar 4. 9 Jumlah Data Yang Terduplikasi

4.2.4.3 *Feature Scaling*

Feature scaling digunakan untuk mengubah variabel asli menjadi variabel bayangan yang dilakukan selama proses klasifikasi (Dwi Nugraheny, 2022). *Feature Scaling* dalam penelitian ini menggunakan metode *standard scaler*. *Standard scaler* digunakan dalam mengantisipasi adanya bias, dimana dilakukan normalisasi untuk melakukan standarisasi pada data untuk menyamakan nilai sehingga tidak ada nilai dengan skala yang beda (Luthfiyah Amatullah, 2022). *Feature scaling* dalam penelitian ini dilakukan untuk melakukan normalisasi pada *dataset*, dimana metode normalisasi yang digunakan *standard scaler*. Pada saat melakukan normalisasi dataset yang digunakan dilakukan pemisahan menjadi 2 variabel yang bertujuan untuk memudahkan dalam melakukan *training data* dan *testing data*. Berikut merupakan hasil normaliasi dengan menggunakan *feature scaling*:

```
[[ 2.  0.  0. ...  0. 30. 16.]
 [ 4.  0.  1. ...  0. 30.  0.]
 [ 4.  1.  0. ...  4. 12.  3.]
 ...
 [ 0.  1.  0. ...  4. 40.  8.]
 [ 4.  1.  0. ...  3. 30. 12.]
 [ 4.  1.  0. ...  1.  5. 12.]]
```

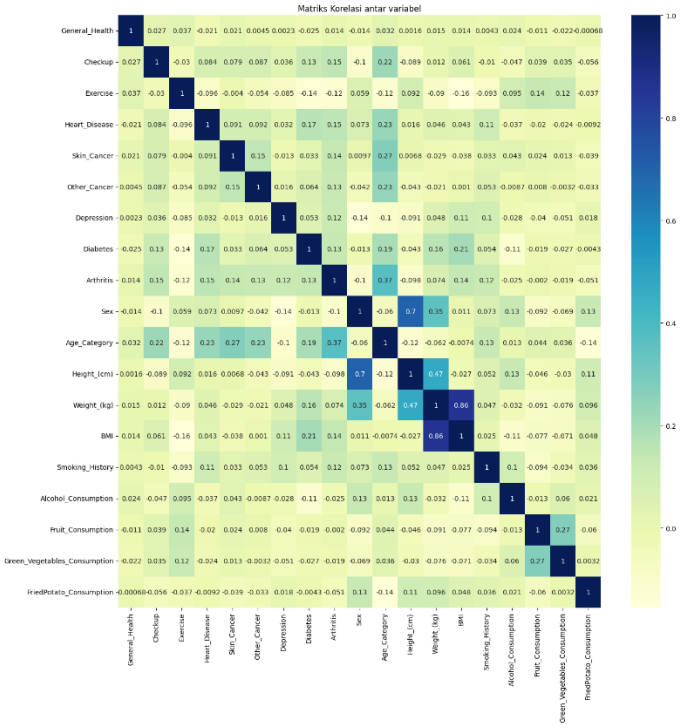
Gambar 4. 10 *Feature Scaling* Variabel X

```
[[12.]
 [ 4.]
 [16.]
 ...
 [ 4.]
 [ 0.]
 [ 1.]]
```

Gambar 4. 11 *Feature Scaling* Variabel Y

4.2.5 Matriks Korelasi

Matriks Korelasi atau *Correlation Matrix* meggambarkan analisis pada setiap matriks memiliki ikatan saling ketergantungan pada atribut dengan dasar yang sama walaupun kadang ditemukan atribut dengan status yang sama (M. Azwan, 2021). *Heatmap* membantu dalam menemukan fitur yang memiliki keterkaitan pada variabel target (Ishak, 2022). Matriks korelasi atau *correlation matrix* dalam penelitian ini digunakan untuk mengukur hubungan antar variabel dengan visualisasi *heatmap* untuk mengetahui variabel yang memiliki hubungan kedekatan antar variabel. Berikut ini merupakan hasil visualisaasi matriks korelasi:



Gambar 4. 12 Visualisasi Matriks Korelasi

4.2.6 Deskripsi Data

Deskripsi data atau *data describe* adalah salah satu fungsi dalam *library pandas* yang digunakan dalam memberikan detail statistik dasar pada nilai numerik (Bhutani, 2023). *Data Describe* atau data deskripsi memberikan gambaran mengenai isi ringkasan dari deskripsi data yang disajikan. Deskripsi data tersebut ditampilkan dengan *transpose* untuk memudahkan dalam membaca deskripsi data. Berikut merupakan hasil dari deskripsi data pada *dataset* diatas:

	count	mean	std	min	25%	50%	75%	max
General_Health	308774.0	2.273232	1.494016	0.00	1.00	2.00	4.00	4.00
Checkup	308774.0	3.514975	1.019649	0.00	4.00	4.00	4.00	4.00
Exercise	308774.0	0.775017	0.417572	0.00	1.00	1.00	1.00	1.00
Heart_Disease	308774.0	0.080871	0.272638	0.00	0.00	0.00	0.00	1.00
Skin_Cancer	308774.0	0.097133	0.296139	0.00	0.00	0.00	0.00	1.00
Other_Cancer	308774.0	0.096760	0.295631	0.00	0.00	0.00	0.00	1.00
Depression	308774.0	0.200467	0.400350	0.00	0.00	0.00	0.00	1.00
Diabetes	308774.0	0.308232	0.724454	0.00	0.00	0.00	0.00	3.00
Arthritis	308774.0	0.327304	0.469230	0.00	0.00	0.00	1.00	1.00
Sex	308774.0	0.481320	0.499652	0.00	0.00	0.00	1.00	1.00
Age_Category	308774.0	6.536104	3.523495	0.00	4.00	7.00	9.00	12.00
Height_(cm)	308774.0	170.615220	10.658452	91.00	163.00	170.00	178.00	241.00
Weight_(kg)	308774.0	83.590399	21.344664	24.95	68.04	81.65	95.25	293.02
BMI	308774.0	28.626813	6.522810	12.02	24.21	27.44	31.85	99.33
Smoking_History	308774.0	0.405662	0.491021	0.00	0.00	0.00	1.00	1.00
Alcohol_Consumption	308774.0	5.097557	8.200434	0.00	0.00	1.00	6.00	30.00
Fruit_Consumption	308774.0	29.834290	24.877812	0.00	12.00	30.00	30.00	120.00
Green_Vegetables_Consumption	308774.0	15.109517	14.926912	0.00	4.00	12.00	20.00	128.00
FriedPotato_Consumption	308774.0	6.297237	8.583837	0.00	2.00	4.00	8.00	128.00

Gambar 4. 13 Hasil Deskripsi Data

4.2.7 *Data Training* dan *Data Testing*

Data training yaitu kelompok data dengan label atau kelas untuk mengidentifikasi karakteristik kelompok data dengan mesin untuk mendapatkan suatu pola atau model data (Wilem Musu, 2021)**.** *Data training* digunakan bertujuan untuk melihat dan mempelajari data, kemudian dilakukan prediksi untuk mengetahui hasil yang dapat membantu dalam pengambilan keputusan (Vibhutijain, 2021).

Sedangkan *Data testing* merupakan kelompok data dengan label atau kelas untuk menguji keakuratan pola atau model data yang telah diklasifikasi (Wilem Musu, 2021)**.** *Data testing* digunakan setelah model di eksekusi yang bertujuan untuk mengevaluasi model setelah dilakukan training model (Vibhutijain, 2021).

Dalam melakukan *training* maupun *testing* data dilakukan pemisahan atau *split data* untuk memisahkan antara *data training* dan *data testing*. Kemudian dilakukan penetapan variabel X untuk melakukan pelatihan dan variabel Y untuk kolom yang menjadi target prediksi. Kemudian juga menghapus data yang tidak memiliki hubungan kedekatan antara satu variabel dengan variabel lain berdasarkan hasil visualisasi matriks korelasi. Selanjutnya dalam melakukan *training* dan *testing data* dilakukan dengan menerapkan pembobotan melalui perbandingan 7:3, dimana 70% untuk *training data* dan 30% untuk *testing data*. Kemudian juga menerapkan metode *SMOTE* (*Synthetic Minority Oversampling Technique*) dalam mengatasi nilai yang tidak mengalami ketidaknormalan selama proses pemodelan. SMOTE adalah teknik yang digunakan untuk mengatasi kelas yang mengalami ketidakseimbangan untuk mendapatkan hasil yang baik dan efektif (Amin Nur Rais, 2019).

```
Jumlah data trainig sebanyak: 216141
Jumlah data testing sebanyak: 92633
```

Gambar 4. 14 Jumlah *Data Training* dan *Data Testing* tanpa SMOTE

```
Jumlah data trainig sebanyak: 397324
Jumlah data testing sebanyak: 170282
```

Gambar 4. 15 Jumlah *Data Training* dan *Data Testing* dengan SMOTE

4.3 Pemodelan data dengan *AutoGluon*

Perancangan model dengan *AutoGluon* dilakukan dengan menggunakan *data training*, dimana label yang digunakan dalam penelitian ini yaitu terdapat pada kolom *'Heart_Disease'*. Pemilihan kolom *Heart_Disease* karena *heart disease* atau penyakit jantung bagian dari penyakit kardiovaskular. Kemudian *problem_type* yang digunakan

dalam penelitian ini yaitu *binary*, dimana untuk menyelesaikan permasalahan dalam memprediksi klasifikasi dengan nilai biner. *Eval_metric* dalam model prediksi ini yaitu *accuracy* yang digunakan untuk mengetahui nilai akurasi pada model yang dihasilkan. *Path* digunakan untuk menyimpan hasil prediksi *AutoGluon* ke dalam direktori *AutoGluonModels*. *Train_data = X_train* dalam prediksi menandakan bahwa data yang digunakan dalam melakukan prediksi adalah data training. *Time_limit* yang digunakan dalam prediksi selama 300 detik atau 5 menit yang bertujuan agar waktu prediksi tidak terlalu cepat atau terlalu lama dalam menghasilkan nilai *accuracy*. *Presets* yang digunakan dalam prediksi ini adalah *best_quality*, dimana dalam prediksi ini bertujuan untuk mengetahui hasil prediksi prediktif dengan akurasi terbaik.

Selanjutnya setelah menjalankan pada pemodelan *AutoGluon*, maka selanjutnya membuat ringkasan hasil prediksi dengan perintah *fit_summary* dan juga *leaderboard*. Hal ini bertujuan untuk merekap hasil prediksi secara keseluruhan pada data dengan jenis tabular yang ditulis secara ringkas. Selain itu juga untuk mengetahui hasil pemeringkatan model dengan hasil prediksi yang terbaik.

Kemudian menapilkan hasil visualisasi hasil prediksi dengan menggunakan *barchart* yang bertujuan untuk menampilkan hasil perbandingan antar model prediksi. Visualisasi hasil prediksi ini dilakukan berdsasarkan hasil pemeringkatan yang telah dilakukan sebelumnya.

	model	score_val	eval_metric	pred_time_val	fit_time	pred_time_val_marginal	fit_time_marginal	stack_level	can_infer	fit_order
0	RandomForestGini_BAG_L2	0.939724	accuracy	88.657702	56.108116	2.667743	52.559307	2	True	3
1	WeightedEnsemble_L3	0.939724	accuracy	89.653601	77.460698	0.995899	21.352582	3	True	4
2	KNeighborsUnif_BAG_L1	0.845119	accuracy	85.989959	3.548810	85.989959	3.548810	1	True	1
3	WeightedEnsemble_L2	0.845119	accuracy	86.627817	3.641543	0.637858	0.092733	2	True	2

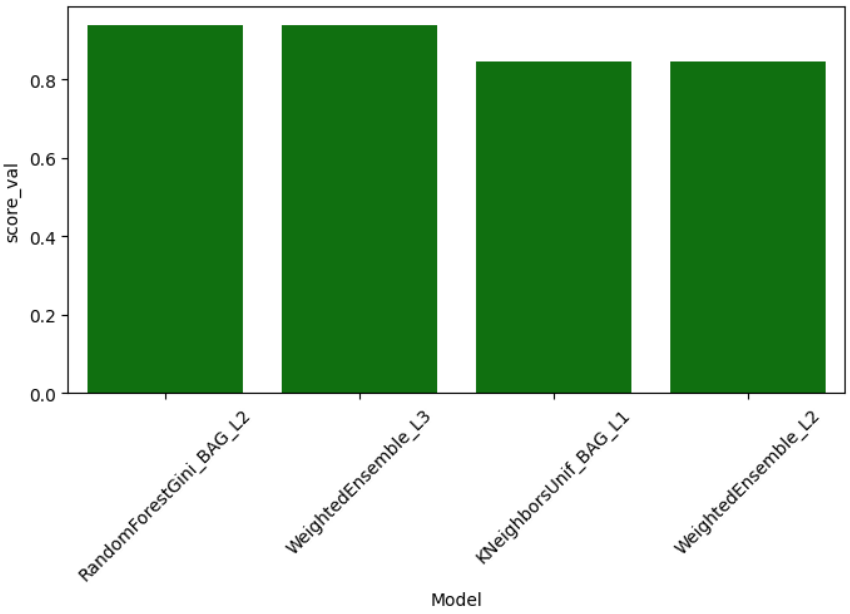
Gambar 4. 16 Hasil Pemodelan *AutoGluon*

```
WeightedEnsemble_L3
```

Gambar 4. 17 Model *AutoGluon* Terbaik

```
{'accuracy': 0.9899250989117194,
 'balanced_accuracy': 0.9899252001954016,
 'mcc': 0.9798502312793361,
 'roc_auc': 0.9994847850508606,
 'f1': 0.9899190349673243,
 'precision': 0.9897769563531432,
 'recall': 0.9900611543770212}
```

Gambar 4. 18 Nilai evaluasi data



Gambar 4. 19 Perbandingan nilai valensi berdasarkan model

```
Nilai Akurasi  Random Forest: 0.93
Hasil Klasifikasi Random Forest:
              precision    recall  f1-score   support

           0       0.93      0.93      0.93     84993
           1       0.93      0.93      0.93     85289

    accuracy                           0.93    170282
   macro avg       0.93      0.93      0.93    170282
weighted avg       0.93      0.93      0.93    170282
```

Gambar 4. 20 Nilai Akurasi model *Random Forest*

```
Nilai Akurasi K-NN: 0.87
Hasil Klasifikasi k-NN:
              precision    recall  f1-score   support

           0       0.99      0.74      0.85     84993
           1       0.79      1.00      0.88     85289

    accuracy                           0.87    170282
   macro avg       0.89      0.87      0.87    170282
weighted avg       0.89      0.87      0.87    170282
```

Gambar 4. 21 Nilai Akurasi model *K-NN*

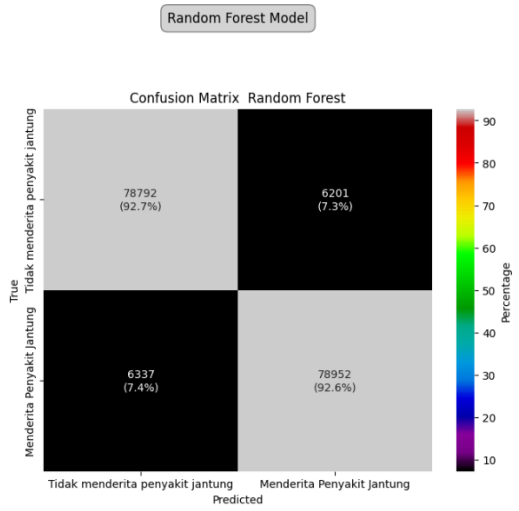
```
Nilai Akurasi Weighted Ensemble Learning: 0.94
Hasil Klasifikasi Weighted Ensemble Learning:
              precision    recall  f1-score   support

           0       0.97      0.91      0.94     84993
           1       0.91      0.97      0.94     85289

    accuracy                           0.94    170282
   macro avg       0.94      0.94      0.94    170282
weighted avg       0.94      0.94      0.94    170282
```

Gambar 4. 22 Nilai Akurasi model *Weighted Ensemble Learning*

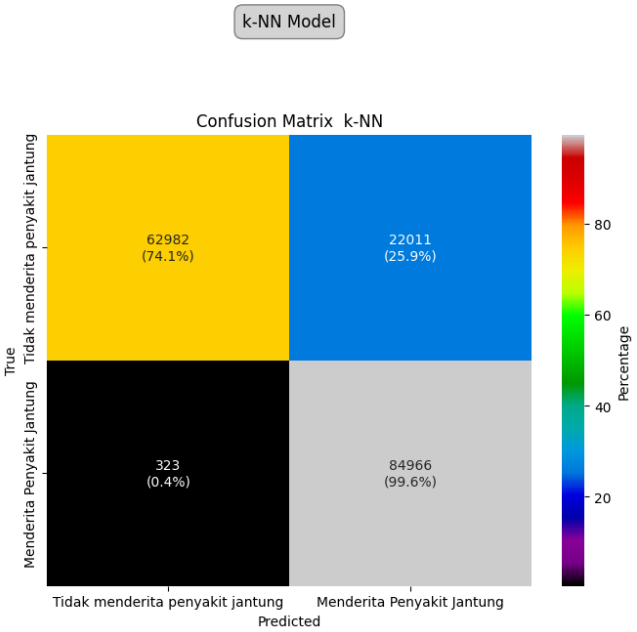
4.4 Evaluasi model

4.4.1 *Confusion Matrix*

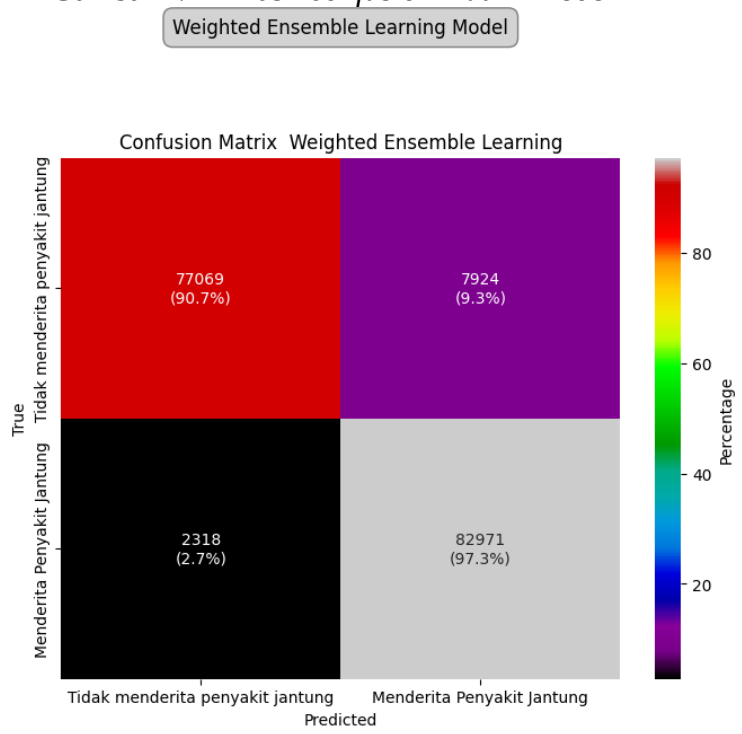
Evaluasi model yang dilakukan dalam penelitian ini menggunakan *confusion matrix* atau matriks konfusi dilakukan dengan melakukan prediksi untuk mengetahui kinerja dari model evaluasi. Kemudian membuat visualisasi *confusion matrix* yang dinormalisasi dengan satuan persen untuk memudahkan dalam membaca matriks. Selain itu juga terdapat *heatmap* memudahkan dalam membaca pola distribusi data mengenai gambaran hasil prediksi yang baik dan tidak.



Gambar 4. 23 Hasil *confusion matrix* model *Random Forest*



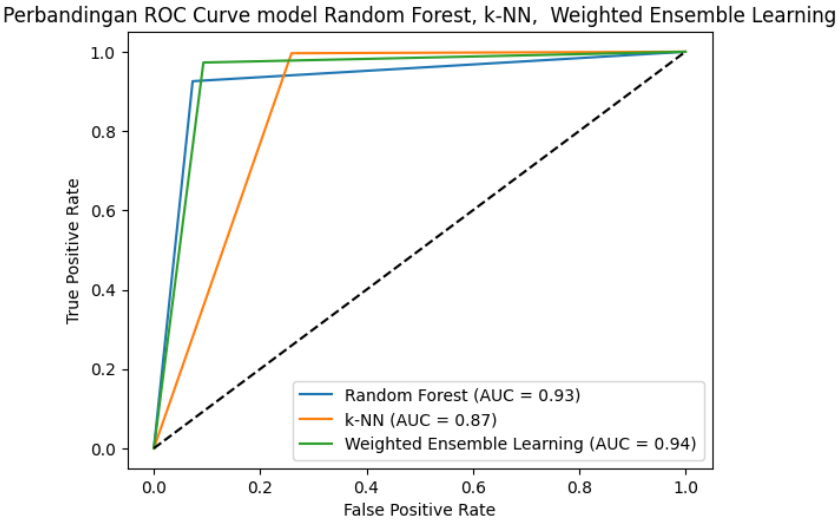
Gambar 4. 24 Hasil *confusion matrix* model *K-NN*



Gambar 4. 25 Hasil *confusion matrix* model *Weighted Ensemble Learning*

4.4.2 ROC Curve

Evaluasi model selanjutnya yang digunakan dalam penelitian ini yaitu *ROC Curve* dilakukan dengan melakukan perbandingan antar model untuk mengetahui perbedaan tingkat sensitivitas kelas positif yang ditandai dengan *true positive rate* (tpr) dan kelas negatif yang ditandai dengan *false positive rate* (fpr). Kemudian *ROC Curve* menyajikan perbandingan model klasifikasi pada area dibawah kurva atau *Area Under Curve* (AUC) yang digunakan dalam melihat kinerja model. Dalam *ROC Curve* terdapat garis diagonal yang menandakan tingkat baik tidaknya model.



Gambar 4. 26 *ROC Curve*

BAB V

HASIL DAN PEMBAHASAN

5.1 Analisis Pre-processing Data

Analisis yang dibahas pada Bab V dilakukan untuk membahas mengenai hasil dari pengumpulan dan pengolah data yang dilakukan pada Bab IV. Pre-*processing data* yang dilakukan dalam penelitian ini bertujuan untuk menyiapkan dan memastikan data agar siap digunakan dalam menjalankan pemodelan dengan menggunakan *AutoGluon*. Dalam menjalankan *pre-processing data* dilakukan dengan menjalankan langkah-langkah untuk menyiapkan data agar sesuai dengan kebutuhan penelitian. Berikut ini merupakan hasil dari langkah *pre-processing data* yang telah dilakukan:

5.1.1 Analisis Informasi Data

Penelusuran informasi dengan perintah *data.info()* yang dilakukan dengan menggunakan *library pandas* pada *dataframe* dilakukan untuk mengumpulkan gambaran mengenai ringkasan yang ada dalam *dataframe* tersebut. Berikut merupakan hasil yang diperoleh melalui perintah *data.info()*:

Tabel 5. 1 Hasil Informasi Data

Jumlah kolom data	**Jumlah baris data**	**Nomor baris awal data**	**Nomor baris akhir data**
19	308.854	0	308.853

Tabel 5. 2 Hasil Informasi Data

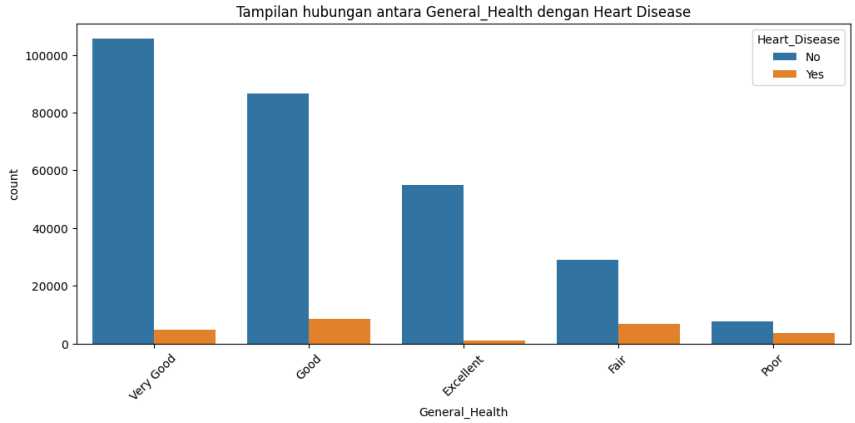
Jenis tipe data	**Jumlah baris *float64***	**Jumlah baris *object***	**Jumlah penggunaan memori (MB)**
Float64	7	12	44.8+
Object			

Informasi dan data yang didapatkan dari fungsi *data.info()* dapat dilihat pada lampiran 19 pada halaman C-8. Kemudian terdapat analisis antara hubungan setiap variabel dengan

variabel *heart disease*. Berikut merupakan analisis dari hubungan antar setiap variabel dengan variabel *heart disease*:

5.1.1.1 Analisis Hubungan Variabel *General Health* dengan Variabel *Heart Disease*

Hubungan variabel *general health* dengan variabel *heart disease* yaitu untuk menjelaskan kondisi kesehatan masyarakat secara umum terhadap penyakit jantung. Secara umum kondisi kesehatan umum masyarakat menggambarkan bagaimana kondisi kesehatan masyarakat saat ini. *Dataset* yang digunakan dalam penelitian ini memberikan gambaran dengan nilai *atribut* yaitu *excellent* menggambarkan kondisi kesehatan masyarakat dengan sangat bagus, kemudiaan *fair* menggambarkan kondisi kesehatan masyarakat sedang dimana kondisi kesehatan masyarakat tidak baik maupun tidak buruk, *good* menggambarkan kondisi kesehatan masyarakat yang baik, *poor* menggambarkan kondisi masyarakat yang buruk, dan *very good* menggambarkan kondisi kesehatan masyarakat yang relatif baik. Kondisi kesehatan masyarakat secara umum ini nantinya akan dibandingkan dengan variabel *heart disease* dengan atribut *yes* dan *no*. Hal ini bertujuan untuk mengetahui sejauh man kondisi Kesehatan masyarakat yang menderita penyakit jantung. Berikut ini merupakan gambar yang menggambarkan hubungan Variabel *General Health* dengan Variabel *Heart Disease*:



Gambar 5. 1 Tampilan Hubungan antara Variabel *General Health* dengan Variabel *Heart Disease*

\Jumlah Variabel General_Health yang berhubungan dengan Variabel Heart Disease:

	General_Health	Heart_Disease	Jumlah	Persentase
0	Excellent	No	54839	98.007292
1	Excellent	Yes	1115	1.992708
2	Fair	No	29021	81.041608
3	Fair	Yes	6789	18.958392
4	Good	No	86721	90.936832
5	Good	Yes	8643	9.063168
6	Poor	No	7729	68.211102
7	Poor	Yes	3602	31.788898
8	Very Good	No	105573	95.632049
9	Very Good	Yes	4822	4.367951

Gambar 5. 2 Jumlah atribut dalam Hubungan antara Variabel *General Health* dengan Variabel *Heart Disease*

Berdasarkan Gambar 5.1 dan Gambar 5.2, dapat diketahui nilai tertinggi pada masing-masing atribut yaitu *excellent* yang tidak terkena penyakit jantung memiliki jumlah sebanyak 54.839 orang dengan persentase sebanyak 98%. Kemudian pada atribut *fair* yang tidak terkena penyakit jantung memiliki jumlah sebanyak 29.021 orang dengan persentase sebanyak 81%. Selanjutnya pada atribut *good* yang tidak terkena penyakit jantung memiliki jumlah sebanyak 86.721 orang dengan persentase sebanyak 91%. Lalu pada atribut *poor* yang tidak terkena penyakit jantung memiliki jumlah sebanyak 7.729 orang dengan persentase sebanyak 68%. Pada atribut *very good* yang tidak terkena penyakit jantung memiliki jumlah sebanyak 105.573 orang dengan persentase sebanyak 96%. Penggunaan perhitungan dengan persentase bertujuan untuk membantu dalam memahami gambaran representasi data. Selain itu juga untuk memudahkan dalam membandingkan antara kondisi kesehatan yang memiliki penyakit jantung dengan kondisi kesehatan yang tidak memiliki penyakit jantung.

Kemudian jumlah masing-masing atribut dengan penyakit jantung yaitu pada atribut *excellent* memiliki jumlah sebanyak 1.115 orang dengan persentase 2%. Selanjutnya pada atribut *fair* memiliki jumlah sebanyak 6.789 orang dengan persentase 19%, pada atribut *good* memiliki jumlah sebanyak 8.643 orang dengan persentase 9%. Lalu pada atribut *poor* memiliki jumlah sebanyak 3.602 orang dengan persentase 32% dan pada atribut *very good* memiliki jumlah sebanyak 4.822 orang dengan persentase 4%.

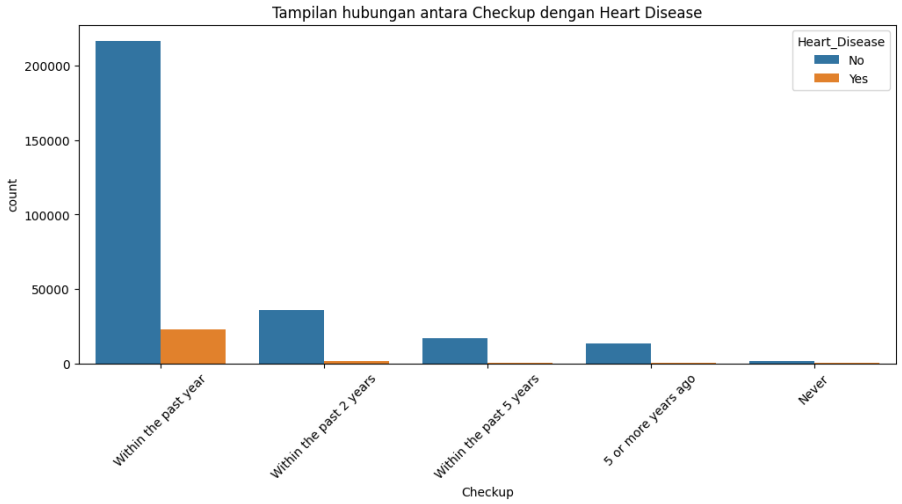
Berdasarkan hasil data tersebut dapat diketahui bahwa masyarakat dengan kondisi kesehatan secara umum tidak terkena penyakit jantung, dimana proporsi persentase kondisi kesehatan tertinggi terdapat pada kondisi kesehatan *excellent* dengan persentase 98% dan jumlah masyarakat yang tidak terkena penyakit jantung tertinggi pada kondisi

kesehatan *very good* dengan jumlah sebanyak 105.573 orang menandakan jumlah. Nilai persentase dan jumlah pada kondisi kesehatan yang berbeda terjadi karena adanya perbedaan dalam persebaran distribusi data pada masing-masing atribut.

Kemudian nilai tertinggi pada proporsi persentase terdapat kondisi kesehatan terkena penyakit jantung terdapat pada kondisi *poor* dengan persentase 32% dan jumlah masyarakat yang terkena penyakit jantung terdapat pada kondisi *good* dengan jumlah sebanyak 8.643 orang. Kondisi kesehatan *poor* lebih rentan terkena penyakit jantung karena persentase proporsi kondisi kesehatan yang tinggi dibandingkan dengan kondisi kesehatan pada atribut yang lainnya walaupun dengan jumlah yang rendah. Sehingga kondisi kesehatan *poor* menjadi kondisi kesehatan yang buruk dan rentan terkena penyakit jantung yang menjadi perhatian khusus dalam penanganan penyakit jantung.

5.1.1.2 Analisis Hubungan Variabel *Checkup* dengan Variabel *Heart Disease*

Hubungan variabel *checkup* dengan variabel *heart disease* yaitu untuk menjelaskan kegiatan yang dilakukan oleh masyarakat untuk memeriksa kesehatan terkait penyakit jantung. Kegiatan pemeriksaan kesehataan menggambarkan kebiasaan masyarakat untuk memeriksa kesehatan secara rutin khususnya terhadap penyakit jantung. Atribut yang terdapat dalam *dataset* yang terdapat dalam penelitian ini yaitu *5 or more years ago* menggambarkan waktu masyarakat pemeriksaan kesehatan selama 5 tahun atau lebih, kemudian *never* menggambarkan masyarakat tidak pernah melakukan kesehatan rutin, *within the past 2 years* menggambarkan masyarakat melakukan pemeriksaan selama 2 tahun terakhir, *within the past 5 years* menggambarkan masyarakat melakukan pemeriksaan selama 5 tahun terakhir, *within the past year* menggambarkan masyarakat melakukan pemeriksaan selama setahun terakhir. Hal ini bertujuan untuk mengtahui seberapa lama masyarakat melakukan pemeriksaan terhadap penyakit jantung. Berikut ini merupakan gambar yang menggambarkan hubungan Variabel *Checkup* dengan Variabel *Heart Disease*:



Gambar 5. 3 Tampilan Hubungan antara Variabel *Checkup* dengan Variabel *Heart Disease*

\Jumlah Variabel Checkup yang berhubungan dengan Variabel Heart Disease:

	Checkup	Heart_Disease	Jumlah	Persentase
0	5 or more years ago	No	13079	97.451755
1	5 or more years ago	Yes	342	2.548245
2	Never	No	1349	95.877754
3	Never	Yes	58	4.122246
4	Within the past 2 years	No	35748	96.063204
5	Within the past 2 years	Yes	1465	3.936796
6	Within the past 5 years	No	16971	97.299622
7	Within the past 5 years	Yes	471	2.700378
8	Within the past year	No	216736	90.543967
9	Within the past year	Yes	22635	9.456033

Gambar 5. 4 Jumlah atribut dalam Hubungan antara Variabel *Checkup* dengan Variabel *Heart Disease*

Berdasarkan Gambar 5.3 dan Gambar 5.4, dapat diketahui nilai tertinggi pada masing-masing atribut yaitu *5 or more years ago* yang tidak terkena penyakit jantung yang memiliki jumlah sebanyak 13.079 orang dengan persentase sebanyak 97%. Kemudian pada atribut *never* yang tidak terkena penyakit jantung yang memiliki jumlah sebanyak 1.349 orang dengan persentase sebanyak 96%. Selanjutnya pada atribut *within the past 2 years* yang tidak akan penyakit jantung yang memiliki jumlah sebanyak 35.748 orang dengan persentase sebanyak 96%. Lalu pada atribut *within the past 5 years* memiliki jumlah sebanyak 16.971 orang dengan persentase sebanyak 97%. Pada atribut *within the past year* yang memiliki jumlah sebanyak 216.736 orang dengan persentase sebanyak 91%. Penggunaan perhitungan dengan persentase bertujuan untuk membantu dalam memahami gambaran representasi data. Selain itu juga untuk memudahkan dalam

membandingkan antara kondisi kesehatan yang memiliki penyakit jantung dengan kondisi kesehatan yang tidak memiliki penyakit jantung. *within the past 5 year*

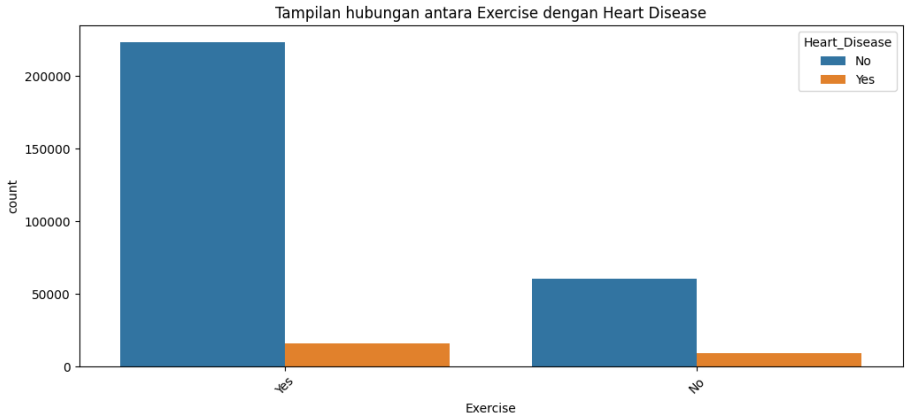
Kemudian jumlah masing-masing atribut dengan penyakit jantung yaitu pada atribut *5 or more years ago* yang memiliki jumlah sebanyak 342 orang dengan persentase sebanyak 3%. Kemudian pada atribut *never* yang memiliki jumlah sebanyak 58 orang dengan persentase sebanyak 4%. Selanjutnya pada atribut *within the past 2 years* yang memiliki jumlah sebanyak 1.465 orang dengan persentase sebanyak 4%. Lalu pada atribut *within the past 5 years* memiliki jumlah sebanyak 471 orang dengan persentase sebanyak 3%. Pada atribut *within the past year* yang memiliki jumlah sebanyak 22.365 orang dengan persentase sebanyak 9%. *within the past 5 year*

Berdasarkan hasil data tersebut dapat diketahui bahwa *checkup* atau pemeriksaan kesehatan yang dilakukan oleh masyarakat mayoritas tidak terkena penyakit jantung dengan proporsi persentase pemeriksaan Kesehatan tertinggi terdapat pada atribut *5 or more years ago* yaitu 97%. Kemudian jumlah masyarakat yang melakukan pemeriksaan kesehatan dan tidak terkena penyakit jantung sebanyak 216.736 orang pada atribut *within the past year*. Nilai persentase dan jumlah pada kondisi kesehatan yang berbeda terjadi karena adanya perbedaan dalam persebaran distribusi data pada masing-masing atribut.

Kemudian nilai tertinggi pada proporsi persentase terdapat *checkup* atau pemeriksaan kesehatan yang terkena penyakit jantung terdapat pada *within the past year* dengan persentase 9% dan jumlah masyarakat yang terkena penyakit jantung tetapi melakukan pemeriksaan kesehatan juga terdapat pada atribut *within the past year* dengan jumlah sebanyak 22.635 orang. Masyarakat yang melakukan pemeriksaan kesehatan rutin setiap tahunnya menjadi kelompok mayoritas dengan jumlah masyarakat yang tidak terkena penyakit jantung terbanyak. Selain itu juga *within the past year* menjadi kelompok masyarakat dengan proporsi persebaran dan jumlah penyakit jantung terbanyak. Hal ini menandakan persebaran distribusi data terpusat pada atribut *within the past year* yang juga menandakan bahwa masyarakat rutin melakukan pemeriksaan kesehatan setiap tahunnya.

5.1.1.3 Analisis Hubungan Variabel *Exercise* dengan Variabel *Heart Disease*

Hubungan variabel *exercise* dengan variabel *heart disease* yaitu untuk menjelaskan kegiatan aktivitas fisik yang dilakukan masyarakat terhadap masyarakat menderita penyakit jantung. Aktivitas fisik menggambarkan kegiatan yang dilakukan oleh masyarakat untuk menjaga kesehatan tubuh masyarakat. Atribut dalam *dataset* yang digunakan terdiri dari *yes* artinya melakukan kegiatan fisik dan *no* artinya tidak melakukan kegiatan fisik. Hal ini bertujuan untuk mengetahui aktivitas fisik yang dilakukan dapat mempengaruhi masyarakat menderita penyakit jantung. Berikut ini merupakan gambar yang menggambarkan hubungan variabel *exercise* dengan variabel *heart disease*:



Gambar 5. 5 Tampilan Hubungan antara Variabel *Exercise* dengan Variabel *Heart Disease*

```
\Jumlah Variabel Exercise yang berhubungan dengan Variabel Heart Disease:
  Exercise Heart_Disease  Jumlah  Persentase
0       No            No   60469   87.039569
1       No           Yes    9004   12.960431
2      Yes            No  223414   93.329880
3      Yes           Yes   15967    6.670120
```

Gambar 5. 6 Jumlah atribut dalam Hubungan antara Variabel *Exercise* dengan Variabel *Heart Disease*

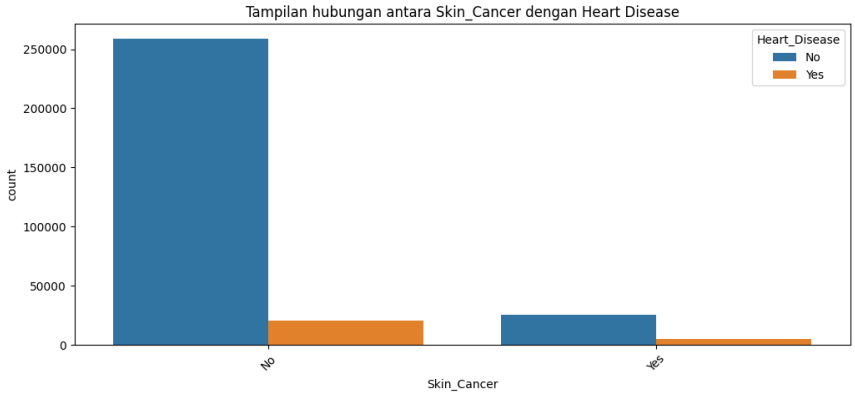
Berdasarkan Gambar 5.5 dan Gambar 5.6, dapat diketahui bahwa nilai tertinggi pada masing-masing atribut yaitu *yes* yang tidak terkena penyakit jantung dengan jumlah sebanyak 223.414 orang dengan persentase sebesar 93%. Kemudian pada atribut *no* yang tidak terkena penyakit jantung dengan jumlah sebanyak 60.469 dengan persentase sebesar 87%. Selanjutnya atribut yang terkena penyakit jantung yaitu *yes* dengan jumlah 15.967

orang dengan persentase 7%. Lalu pada atribut *no* dengan jumlah 20.281 orang dengan persentase sebesar 13%.

Berdasarkan data tersebut dapat diketahui bahwa mayoritas masyarakat yang melakukan *exercise* atau latihan tidak terkena penyakit jantung, dimana dalam hal ini proporsi persentase pada *exercise* tertinggi dengan nilai 93% dan juga dengan jumlah masyarakat tertinggi yaitu 223.414 orang. Kemudian nilai tertinggi pada proporsi persentase melakukan Latihan tetapi terkena penyakit jantung terdapat pada atribut *no* dengan persentase 13% dan juga jumlah masyarakat tertinggi pada penyakit jantung pada atribut *yes* dengan jumlah 15.967 orang. Nilai persentase dan jumlah pada kondisi kesehatan yang berbeda terjadi karena adanya perbedaan dalam persebaran distribusi data pada masing-masing atribut. Kemudian mayoritas masyarakat telah melakukan *exercise* atau latihan fisik, akan tetapi masih terdapat risiko penyakit jantung walaupun telah melakukan latihan fisik yang dapat disebabkan oleh berbagai faktor. Selanjutnya masyarakat yang tidak melakukan latihan fisik lebih rentan terkena penyakit jantung yang menjadi perhatian khusus. Hal ini dapat dilihat melalui persentase proporsi masyarakat yang tidak melakukan latihan fisik, dimana persentase masyarakat yang tidak melakukan latihan fisik dan terkena penyakit jantung mencapai 32%.

5.1.1.4 Analisis Hubungan Variabel *Skin Cancer* dengan Variabel *Heart Disease*

Hubungan variabel *skin cancer* dengan variabel *heart disease* yaitu untuk menjelaskan masyarakat yang menderita penyakit kanker kulit dengan masyarakat yang menderita penyakit jantung. Kanker kulit menggambarkan masyarakat yang menderita penyakit kanker kulit. Atribut dalam *dataset* yang digunakan terdiri dari *yes* artinya masyarakat menderita penyakit kanker kulit melakukan kegiatan fisik dan *no* artinya masyarakat tidak menderita penyakit kanker kulit. Hal ini bertujuan untuk mengetahui masyarakat yang menderita penyakit kanker kulit dengan penyakit jantung. Berikut ini merupakan gambar yang menggambarkan hubungan variabel *skin cancer* dengan variabel *heart disease*:



Gambar 5. 7 Tampilan Hubungan antara Variabel *Skin Cancer* dengan Variabel *Heart Disease*

\Jumlah Variabel Skin_Cancer yang berhubungan dengan Variabel Heart Disease:

	Skin_Cancer	Heart_Disease	Jumlah	Persentase
0	No	No	258579	92.727175
1	No	Yes	20281	7.272825
2	Yes	No	25304	84.363539
3	Yes	Yes	4690	15.636461

Gambar 5. 8 Jumlah atribut dalam Hubungan antara Variabel *Skin Cancer* dengan Variabel *Heart Disease*

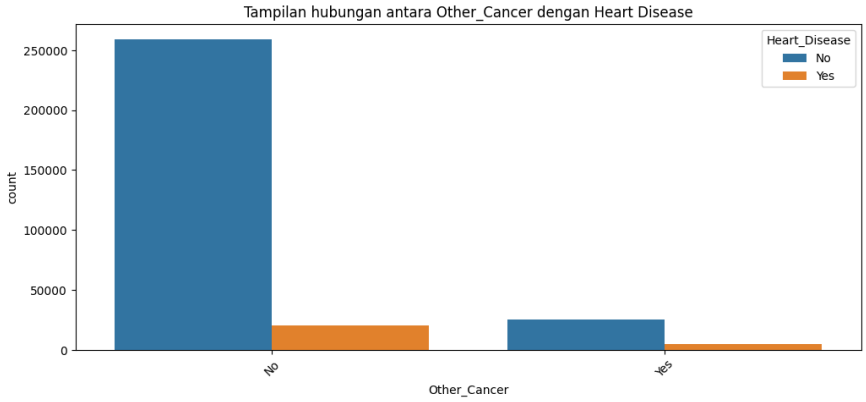
Berdasarkan Gambar 5.7 dan Gambar 5.8, dapat diketahui nilai tertinggi pada masing-masing atribut terdapat pada masyarakat yang tidak terkena penyakit jantung yaitu pada atribut *no* yang memiliki jumlah sebanyak 258.579 orang dengan proporsi persentase sebesar 93%. Selanjutnya pada atribut *yes* memiliki jumlah sebanyak 25.304 dengan proporsi persentase sebesar 84%. Lalu terdapat atribut yang terkena penyakit jantung yaitu atribut *no* dengan jumlah sebanyak 20.281 orang dengan proporsi persentase sebesar 7%. Kemudian atribut *yes* yang terkena penyakit jantung memiliki jumlah sebanyak 4.690 orang dengan persentase 16%.

Berdasarkan data tersebut dapat diketahui bahwa mayoritas masyarakat dalam kondisi tidak terkena penyakit *skin cancer* atau kanker kulit juga tidak menderita penyakit jantung. Dalam hal ini proporsi persentase masyarakat yang tidak terkena penyakit kanker kulit dan tidak penyakit jantung dengan nilai 93% dan berjumlah 258.579 orang. Kemudian jumlah masyarakat yang terkena penyakit kanker kulit dan juga penyakit jantung terdapat pada atribut *no* dengan jumlah 20.256 orang serta proporsi persentase masyarakat yang terkena penyakit kanker kulit dan jantung terdapat pada atribut *yes* dengan persentase sebesar 16%. Nilai jumlah dan proporsi persentase yang berbeda dapat

terjadi karena perbedaan dalam pola persebaran data. Kemudian masyarakat yang tidak terkena penyakit kanker kulit masih terdapat risiko penyakit jantung yang dapat disebabkan oleh berbagai faktor. Selanjutnya masyarakat yang terkena penyakit kanker kulit dan penyakit jantung menjadi perhatian khusus. Hal ini dapat diketahui melalui persentase proporsi masyarakat yang terkena penyakit kanker kulit dan penyakit jantung, dimana persentase masyarakat yang terkena penyakit kanker kulit dan penyakit jantung memiliki nilai mencapai 16% yang dibandingkan dengan masyarakat yang tidak terkena penyakit kanker kulit tetapi terkena penyakit jantung mencapai 7%.

5.1.1.5 Analisis Hubungan Variabel *Other Cancer* dengan Variabel *Heart Disease*

Hubungan variabel *other cancer* dengan variabel *heart disease* yaitu untuk menjelaskan masyarakat yang menderita penyakit kanker dengan jenis lain dengan masyarakat yang menderita penyakit jantung. Kanker dengan jenis yang lain menggambarkan masyarakat menderita penyakit kanker dengan jenis yang lain selain kanker kulit. Atribut dalam *dataset* yang digunakan terdiri dari *yes* artinya masyarakat menderita penyakit kanker dengan jenis lain dan *no* artinya masyarakat tidak menderita penyakit kanker dengan jenis lain. Hal ini bertujuan untuk mengetahui masyarakat yang menderita penyakit kanker jenis yang lain selain kanker kulit dengan penyakit jantung. Berikut ini merupakan gambar yang menggambarkan hubungan variabel *other cancer* dengan variabel *heart disease*:



Gambar 5. 9 Tampilan Hubungan antara Variabel *Other Cancer* dengan Variabel *Heart Disease*

```
\Jumlah Variabel Other_Cancer yang berhubungan dengan Variabel Heart Disease:
  Other_Cancer Heart_Disease  Jumlah  Persentase
0           No            No  258720   92.739160
1           No           Yes   20256    7.260840
2          Yes            No   25163   84.219158
3          Yes           Yes    4715   15.780842
```

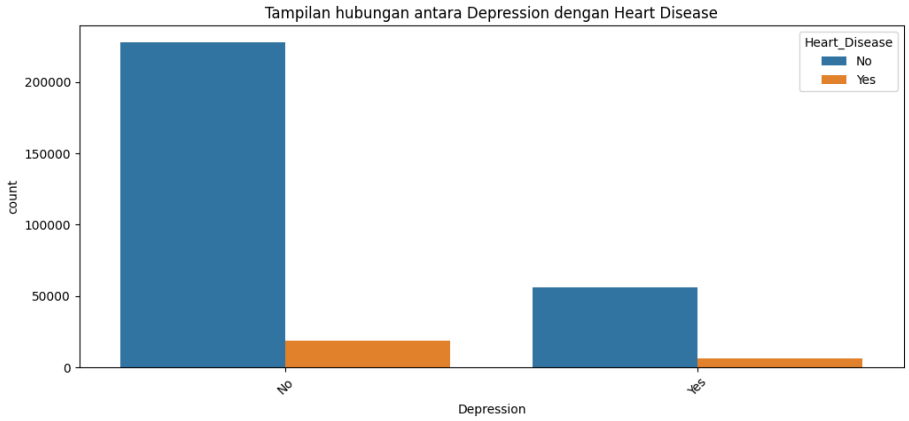
Gambar 5. 10 Jumlah atribut dalam Hubungan antara Variabel *Other Cancer* dengan Variabel *Heart Disease*

Berdasarkan Gambar 5.9 dan Gambar 5.10, dapat diketahui nilai tertinggi pada masing-masing atribut terdapat pada masyarakat yang tidak terkena penyakit jantung yaitu pada atribut *no* yang memiliki jumlah sebanyak 258.720 orang dengan proporsi persentase sebesar 93%. Selanjutnya pada atribut *yes* memiliki jumlah sebanyak 25.163 dengan proporsi persentase sebesar 84%. Lalu terdapat atribut yang terkena penyakit jantung yaitu atribut *no* dengan jumlah sebanyak 20.256 orang dengan proporsi persentase sebesar 7%. Kemudian atribut *yes* yang terkena penyakit jantung memiliki jumlah sebanyak 4.715 orang dengan persentase 16%.

Berdasarkan data tersebut dapat diketahui bahwa mayoritas masyarakat dalam kondisi tidak terkena penyakit *other cancer* atau kanker jenis yang lain juga tidak menderita penyakit jantung. Dalam hal ini proporsi persentase masyarakat yang tidak terkena penyakit kanker kulit dan tidak penyakit jantung dengan nilai 93% dan berjumlah 258.579 orang. Kemudian jumlah masyarakat yang terkena penyakit kanker kulit dan juga penyakit jantung terdapat pada atribut *no* dengan jumlah 20.256 orang serta proporsi persentase masyarakat yang terkena penyakit kanker kulit dan jantung terdapat pada atribut *yes* dengan persentase sebesar 16%. Nilai jumlah dan proporsi persentase yang berbeda dapat terjadi karena perbedaan dalam pola persebaran data. Kemudian masyarakat yang tidak terkena penyakit kanker dengan jenis yang lain masih terdapat risiko penyakit jantung yang dapat disebabkan oleh berbagai faktor. Selanjutnya masyarakat yang terkena penyakit kanker dengan jenis lain dan penyakit jantung menjadi perhatian khusus. Hal ini dapat diketahui melalui persentase proporsi masyarakat yang terkena penyakit kanker dengan jenis lain dan penyakit jantung, dimana persentase masyarakat yang terkena penyakit kanker dengan jenis lain dan penyakit jantung memiliki nilai mencapai 16% yang dibandingkan dengan masyarakat yang tidak terkena penyakit kanker dengan jenis lain tetapi terkena penyakit jantung mencapai 7%.

5.1.1.6 Analisis Hubungan Variabel *Depression* dengan Variabel *Heart Disease*

Hubungan variabel *depression* dengan variabel *heart disease* yaitu untuk menjelaskan masyarakat yang menderita penyakit derpresi terhadap masyarakat yang menderita penyakit jantung. Depresi menggambarkan masyarakat menderita penyakit depresi berupa gangguan mental dengan masyarakat yang menderita penyakit jantung. Atribut yang terdapat dalam *dataset* yang digunakan terdiri dari *yes* artinya masyarakat menderita penyakit depresi dan *no* artinya masyarakat tidak menderita penyakit depresi. Hal ini bertujuan untuk mengetahui masyarakat yang menderita penyakit depresi dengan penyakit jantung. Berikut ini merupakan gambar yang menggambarkan hubungan variabel *depression* dengan variabel *heart disease*:



Gambar 5. 11 Tampilan Hubungan antara Variabel *Depression* dengan Variabel *Heart Disease*

\Jumlah Variabel Depression yang berhubungan dengan Variabel Heart Disease:

	Depression	Heart_Disease	Jumlah	Persentase
0	No	No	228083	92.35887
1	No	Yes	18870	7.64113
2	Yes	No	55800	90.14394
3	Yes	Yes	6101	9.85606

Gambar 5. 12 Jumlah atribut dalam Hubungan antara Variabel *Depression* dengan Variabel *Heart Disease*

Berdasarkan Gambar 5.11 dan Gambar 5.12, dapat diketahui nilai tertinggi pada masing-masing atribut terdapat pada masyarakat yang tidak terkena penyakit jantung yaitu pada atribut *no* yang memiliki jumlah sebanyak 228.083 orang dengan proporsi persentase sebesar 92%. Selanjutnya pada atribut *yes* memiliki jumlah sebanyak 55.800 dengan proporsi persentase sebesar 90%. Lalu terdapat atribut yang terkena penyakit jantung

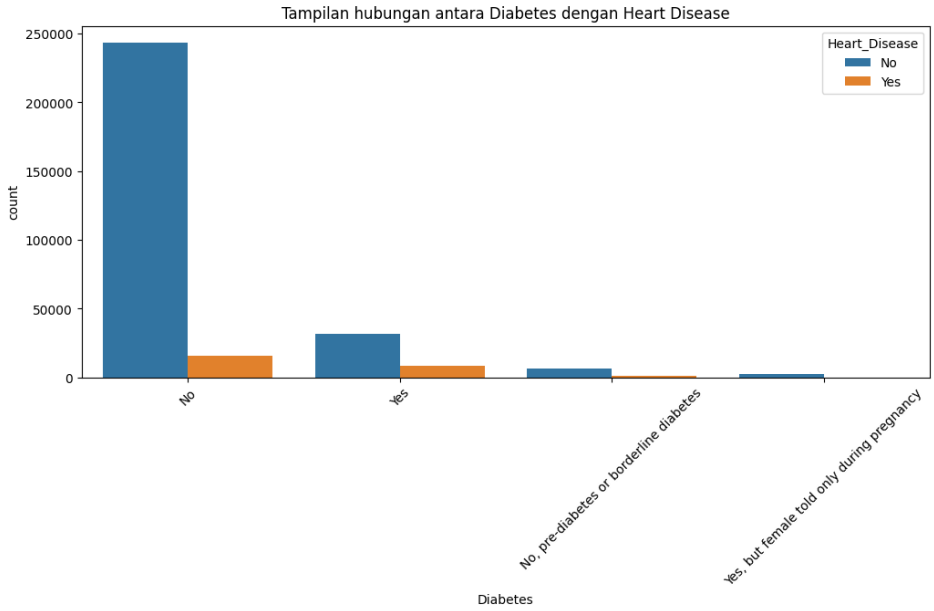
yaitu atribut *no* dengan jumlah sebanyak 18.870 orang dengan proporsi persentase sebesar 8%. Kemudian atribut *yes* yang terkena penyakit jantung memiliki jumlah sebanyak 6.101 orang dengan persentase 10%.

Berdasarkan data tersebut dapat diketahui bahwa mayoritas masyarakat dalam kondisi tidak terkena penyakit depresi juga tidak menderita penyakit jantung. Dalam hal ini proporsi persentase masyarakat yang tidak terkena penyakit depresi dan tidak penyakit jantung dengan nilai 92% dan berjumlah 228.083 orang. Kemudian jumlah masyarakat yang tidak terkena penyakit depresi dan juga penyakit jantung terdapat pada atribut *no* dengan jumlah 18.870 orang serta proporsi persentase masyarakat yang terkena penyakit kanker kulit dan jantung terdapat pada atribut *yes* dengan persentase sebesar 16%. Nilai jumlah dan proporsi persentase yang berbeda dapat terjadi karena perbedaan dalam pola persebaran data. Kemudian masyarakat yang tidak depresi masih terdapat risiko penyakit jantung yang dapat disebabkan oleh berbagai faktor. Selanjutnya masyarakat yang terkena penyakit depresi dan penyakit jantung menjadi perhatian khusus. Hal ini dapat diketahui melalui persentase proporsi masyarakat yang terkena penyakit depresi dan penyakit jantung, dimana persentase masyarakat yang terkena penyakit depresi dan penyakit jantung memiliki nilai mencapai 16% yang dibandingkan dengan masyarakat yang tidak terkena penyakit depresi tetapi terkena penyakit jantung mencapai 7%.

5.1.1.7 Analisis Hubungan Variabel *Diabetes* dengan Variabel *Heart Disease*

Hubungan variabel *diabetes* dengan variabel *heart disease* yaitu untuk menjelaskan masyarakat yang menderita penyakit diabetes terhadap masyarakat yang menderita penyakit jantung. Diabetes menggambarkan masyarakat yang mengalami gula darah yang tinggi. *Dataset* dalam penelitian ini memiliki atribut yaitu *no* artinya masyarakat tidak menderita penyakit diabetes. Kemudian *no, pre-diabetes or borderline diabetes* artinya masyarakat tidak menderita penyakit diabetes tetapi ada *pre-diabetes* atau gula darah dengan kadar yang tinggi tetapi belum mencapai kategori diabetes. Atribut *yes* artinya masyarakat menderita penyakit diabetes. Atribut *yes, but female told only during pregnancy* artinya ya, perempuan menderita penyakit diabetes tetapi diberitahu ketika kondisi selama proses kehamilan. Hal ini bertujuan untuk mengetahui masyarakat yang

menderita penyakit diabetes dengan penyakit jantung. Berikut ini merupakan gambar yang menggambarkan hubungan variabel *diabetes* dengan variabel *heart disease*:



Gambar 5. 13 Tampilan Hubungan antara Variabel *Diabetes* dengan Variabel *Heart Disease*

```
\Jumlah Variabel Diabetes yang berhubungan dengan Variabel Heart Disease:
                                  Diabetes Heart_Disease  Jumlah  \
0                                       No            No  243436
1                                       No           Yes   15705
2     No, pre-diabetes or borderline diabetes           No    6102
3     No, pre-diabetes or borderline diabetes          Yes     794
4                                      Yes            No   31795
5                                      Yes           Yes    8376
6  Yes, but female told only during pregnancy           No    2550
7  Yes, but female told only during pregnancy          Yes      96

   Persentase
0   93.939593
1    6.060407
2   88.486079
3   11.513921
4   79.149137
5   20.850863
6   96.371882
7    3.628118
```

Gambar 5. 14 Jumlah atribut dalam Hubungan antara Variabel *Diabetes* dengan Variabel *Heart Disease*

Berdasarkan Gambar 5.13 dan Gambar 5.14, dapat diketahui nilai tertinggi pada masing-masing atribut antara variabel diabetes dengan variabel heart disease yaitu *no* tidak terkena penyakit jantung dengan jumlah sebanyak 243.436 orang dengan proporsi persentase sebesar 94%. Kemudian pada atribut *yes* tidak terkena penyakit jantung

dengan jumlah sebanyak 31.795 orang dengan proporsi persentase sebesar 79%. Lalu pada atribut *no, pre-diabetes or borderline diabetes* yang tidak terkena penyakit jantung dengan jumlah sebanyak 6.102 orang dengan proporsi persentase sebesar 88%. Selanjutnya pada atribut *yes, but female told only during pregnancy* yang tidak terkena penyakit jantung dengan jumlah sebanyak 2.550 orang dengan proporsi persentase sebesar 96%.

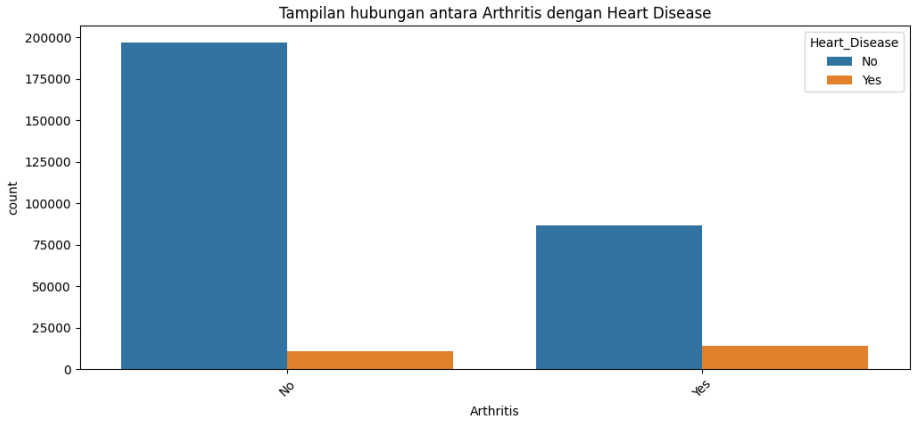
Kemudian jumlah masing-masing atribut dengan penyakit jantung yaitu pada atribut *no* dengan jumlah sebanyak 15.705 orang dengan proprose persentase sebesar 6%. Selanjutnya pada atribut *Yes* memiliki jumlah sebanyak 8.376 orang dengan proporsi persentase 12%. Lalu pada atribut *no, pre-diabetes or borderline diabetes* memiliki jumlah sebanyak 794 orang dengan proporsi persentase 21%. Pada atribut *yes, but female told only during pregnancy* memiliki jumlah sebanyak 94 orang dengan proporsi persentase 4%.

Berdasarkan hasil data tersebut dapat diketahui bahwa mayoritas masyarakat tidak terkena penyakit jantung maupun diabetes, dimana jumlah terbanyak terdapat dalam atribut *no* dengan jumlah sebanyak 243.436 orang dan proporsi persentase terbanyak terdapat dalam atribut *yes, but female told only during pregnancy* dengan persentase sebesar 96%. Jumlah dan proporsi persentase yang berbeda dapat dipengaruhi melalui pola persebaran data yang berbeda. Kemudian masyarakat yang memiliki penyakit jantung dengan jumlah yang tertinggi terdapat dalam atribut *no* dengan jumlah sebanyak 15.705 orang dan persentase tertinggi terdapat dalam atribut *yes* dengan nilai 21%. Sehingga masyarakat yang memiliki penyakit diabetes dan penyakit jantung menjadi perhatian khusus karena rentan terserang terhadap kedua penyakit tersebut.

5.1.1.8 Analisis Hubungan Variabel *Arthritis* dengan Variabel *Heart Disease*

Hubungan variabel *arthritis* dengan variabel *heart disease* yaitu untuk menjelaskan masyarakat yang menderita penyakit arthritis dengan masyarakat yang menderita penyakit jantung. Arthritis menggambarkan masyarakat yang mengalami penyakit arthris berupa radang pada sendi dengan masyarakat yang menderita penyakit jantung. Atribut yang terdapat dalam *dataset* yang digunakan terdiri dari *yes* artinya masyarakat menderita penyakit arthritis dan *no* artinya masyarakat tidak menderita penyakit arthritis. Hal ini

bertujuan untuk mengetahui masyarakat yang menderita penyakit arthritis dengan penyakit jantung. Berikut ini merupakan gambar yang menggambarkan hubungan variabel *arthritis* dengan variabel *heart disease*:



Gambar 5. 15 Tampilan Hubungan antara Variabel *Arthritis* dengan Variabel *Heart Disease*

\Jumlah Variabel Arthritis yang berhubungan dengan Variabel Heart Disease:

	Arthritis	Heart_Disease	Jumlah	Persentase
0	No	No	197064	94.841253
1	No	Yes	10719	5.158747
2	Yes	No	86819	85.899021
3	Yes	Yes	14252	14.100979

Gambar 5. 16 Jumlah atribut dalam Hubungan antara Variabel *Arthritis* dengan Variabel *Heart Disease*

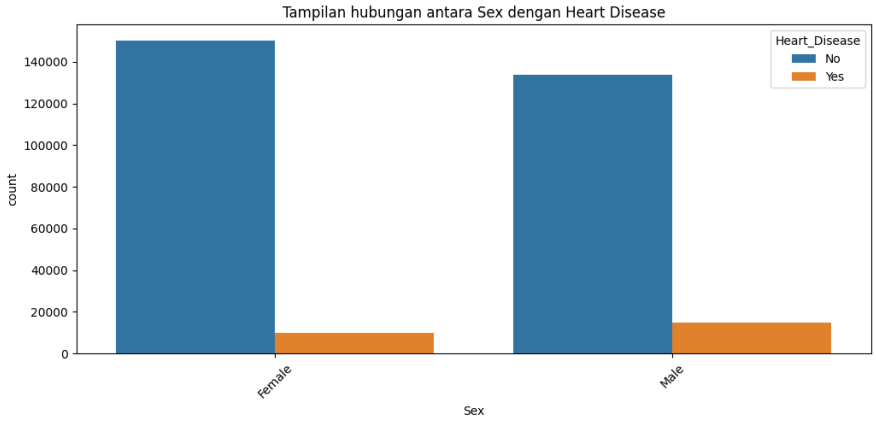
Berdasarkan Gambar 5.15 dan Gambar 5.16, dapat diketahui nilai tertinggi pada masing-masing atribut antara variabel *arthritis* dengan variabel *heart disease* yaitu *no* tidak terkena penyakit jantung dengan jumlah sebanyak 197.064 orang dengan proporsi persentase sebesar 95%. Kemudian pada atribut *yes* tidak terkena penyakit jantung dengan jumlah sebanyak 86.819 orang dengan proporsi persentase sebesar 86%. Kemudian jumlah masing-masing atribut dengan penyakit jantung yaitu pada atribut *no* dengan jumlah sebanyak 10.719 orang dengan proprose persentase sebesar 5%. Selanjutnya pada atribut *Yes* memiliki jumlah sebanyak 14.252 orang dengan proporsi persentase 14%.

Berdasarkan data tersebut dapat diketahui bahwa mayorias masyarakat dalam keadaan tidak menderita penyakit jantung dan juga penyakit arthritis. Dalam hal ini proporsi persentase masyarakat yang tidak terkena penyakit arthritis dan tidak penyakit

jantung dengan nilai 95% dan berjumlah 197.064 orang. Kemudian juga jumlah masyarkat yang menderita penyakit jantung dan juga menderita penyakit arthritis mencapai 14.252 orang dengan proporsi persentese 14%. Sehingga dengan jumlah dan proporsi persentase yang tinggi menjadikan penyakit jantung dan penyakit arthritis sebagai fokus dalam penenganan pencegahan penyakit.

5.1.1.9 Analisis Hubungan Variabel *Sex* dengan Variabel *Heart Disease*

Hubungan variabel *sex* dengan variabel *heart disease* yaitu untuk menjelaskan jenis kelamin masyarakat terhadap masyarakat yang menderita penyakit jantung. Jenis kelamin dalam hal ini menggambarkan jenis kelamin yang mengalami penyakit jantung. Atribut yang terdapat dalam *dataset* yang digunakan terdiri dari *male* artinya masyarakat berjenis kelamin laki-laki dan *female* artinya masyarakat berjenis kelamin perempuan. Hal ini bertujuan untuk mengetahui jenis kelamin masyarakat yang menderita penyakit jantung. Berikut ini merupakan gambar yang menggambarkan hubungan variabel *sex* dengan variabel *heart disease*:



Gambar 5. 17 Tampilan Hubungan antara Variabel *Sex* dengan Variabel *Heart Disease*

```
\Jumlah Variabel Sex yang berhubungan dengan Variabel Heart Disease:
      Sex Heart_Disease  Jumlah  Persentase
0  Female            No  150298   93.821319
1  Female           Yes    9898    6.178681
2    Male            No  133585   89.860620
3    Male           Yes   15073   10.139380
```

Gambar 5. 18 Jumlah atribut dalam Hubungan antara Variabel *Sex* dengan Variabel *Heart Disease*

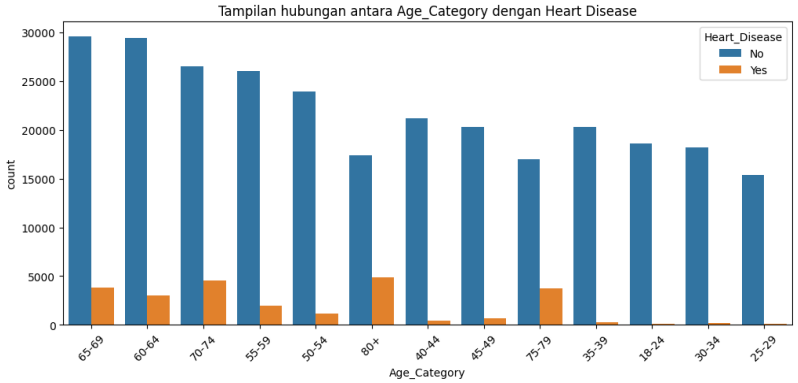
Berdasarkan Gambar 5.17 dan Gambar 5.18, dapat diketahui nilai tertinggi pada masing-masing atribut antara variabel sex dengan variabel heart disease yaitu *female* tidak terkena penyakit jantung dengan jumlah sebanyak 150.298 orang dengan proporsi persentase sebesar 94%. Kemudian pada atribut *male* tidak terkena penyakit jantung dengan jumlah sebanyak 133.585 orang dengan proporsi persentase sebesar 90%. Kemudian jumlah masing-masing atribut dengan penyakit jantung yaitu pada atribut *female* dengan jumlah sebanyak 9.898 orang dengan proprose persentase sebesar 6%. Selanjutnya pada atribut *male* memiliki jumlah sebanyak 15.073 orang dengan proporsi persentase 10%.

Berdasarkan data tersebut dapat diketahui bahwa mayoritas masyarakat memiliki jenis kelamin perempuan dan tidak terkena penyakit jantung. Proporsi persentase masyarakat dengan jenis kelamin perempuan dan tidak penyakit jantung dengan nilai 94% dan berjumlah 150.298 orang. Kemudian jumlah masyarakat dengan jenis kelamin laki-laki dan menderita penyakit jantung memiliki jumlah 15.073 orang dengan proporsi persentase sebesar 10%. Perbedaan distribusi data dalam masing-masing atribut menjadikan nilai persentase dan jumlah yang berbeda. Kemudian masyarakat dengan jenis kelamin laki-laki rentan terhadap penyakit jantung, dimana memiliki nilai persentase dan jumlah penderita yang banyak. Sehingga masyarakat dengan jenis kelamin laki-laki menjadi perhatian khusus dalam menangani penyakit jantung.

5.1.1.10 Analisis Hubungan Variabel *Age Category* dengan Variabel *Heart Disease*

Hubungan variabel *age category* dengan variabel *heart disease* yaitu untuk menjelaskan kategori usia yang menderita penyakit jantung. Kategori usia menggambarkan rentang usia masyarakat yang menderita penyakit jantung. Atribut dalam penelitian ini berupa retang usia yaitu 18-24 tahun, 25-29 tahun, 30-34 tahun, 35-39 tahun, 40-44 tahun, 45-49 tahun, 50-54 tahun, 55-59 tahun, 60-64 tahun, 65-69 tahun, 70-74 tahun, 75-79 tahun, dan usia 80 tahun keatas. Hal ini bertujuan untuk mengetahui rentang usia masyarakat yang terkena penyakit jantung. Pada sub-bab ini disajikan dengan gambar 10 baris teratas, hal ini dikarenakan jumlah baris dalam variabel yang banyak dengan jumlah sebanyak 26 baris. Untuk gambar hubungan variabel *age category* dengan variabel *heart disease* secara keseluruhan dapat dilihat pada gambar di lampiran 1 halaman C-1. Berikut ini

merupakan gambar 10 baris teratas yang menggambarkan hubungan variabel *diabetes* dengan variabel *heart disease*:



Gambar 5. 19 Tampilan Hubungan antara Variabel *Age Category* dengan Variabel *Heart Disease*

Jumlah Variabel Age_Category yang berhubungan dengan Variabel Heart Disease:

	Age_Category	Heart_Disease	Jumlah	Persentase
0	18-24	No	18587	99.496815
1	18-24	Yes	94	0.503185
2	25-29	No	15381	99.270685
3	25-29	Yes	113	0.729315
4	30-34	No	18227	98.909269
5	30-34	Yes	201	1.090731
6	35-39	No	20332	98.670290
7	35-39	Yes	274	1.329710
8	40-44	No	21160	97.985645
9	40-44	Yes	435	2.014355

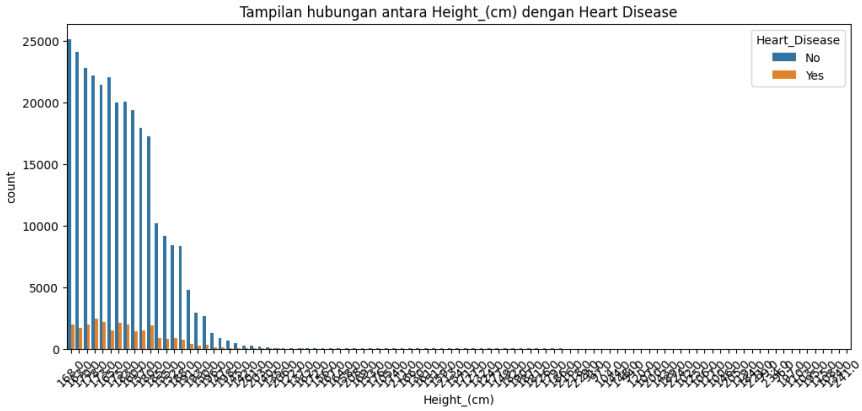
Gambar 5. 20 Jumlah atribut dalam Hubungan antara Variabel *Age Category* dengan Variabel *Heart Disease*

Berdasarkan Gambar 5.19 dan Gambar 5.20, dapat diketahui bahwa kategori usia terbanyak antara variabel *age category* dengan variabel *heart disease* terdapat pada kategori usia 65-69 tahun. Hubungan antara variabel *age category* dan variabel *Heart Disease* terdapat atribut *no* yang menandakan hubungan variabel tersebut tidak terdapat penyakit jantung. Kemudian juga terdapat atribut *yes* yang menandakan hubungan antara kedua variabel tersebut terdapat penyakit jantung. Jumlah masyarakat dalam kategori usia 65-69 tahun mencapai 33.434 orang dimana terdapat orang yang tidak terkena penyakit jantung mencapai 29.611 orang dengan persentase 88% dan orang yang terkena penyakit jantung mencapai 3.823 orang dengan persentase 12%. Selanjutnya pada kategori usia antara 80 keatas terdapat orang yang tidak terkena penyakit jantung mencapai 17.415 orang dengan persentase 78% dan orang yang terkena penyakit jantung mencapai 4.856

orang dengan persentase 22%. Nilai persentase dan jumlah yang berbeda dapat terjadi karena adanya perbedaan dalam pola persebaran data. Sehingga pada kategori usia 80 tahun keatas menjadi perhatian dalam penangan penyakit jantung karena lebih rentan terserang penyakit jantung.

5.1.1.11 Analisis Hubungan Variabel *Height (cm)* dengan Variabel *Heart Disease*

Hubungan variabel *height (cm)* dengan variabel *heart disease* yaitu untuk menjelaskan tinggi badan masyarakat yang menderita penyakit jantung. Atribut yang terdapat pada variabel ini mulai dari 91 cm, 94 cm, 96 cm, 97 cm, hingga 241 cm. Hal ini bertujuan untuk mengetahui tinggi badan masyarkat yang menderita penyakit jantung. Pada sub-bab ini disajikan dengan gambar 10 baris teratas, hal ini dikarenakan jumlah baris dalam variabel yang banyak dengan jumlah sebanyak 151 baris. Hubungan variabel *height (cm)* dengan variabel *heart disease* dapat dilihat pada lampiran 2 pada halaman C-1. Berikut ini merupakan gambar 10 baris teratas yang menggambarkan hubungan variabel *height (cm)* dengan variabel *heart disease*:



Gambar 5. 21 Tampilan Hubungan antara Variabel *Height (cm)* dengan Variabel *Heart Disease*

Jumlah Variabel Height_(cm) yang berhubungan dengan Variabel Heart Disease:

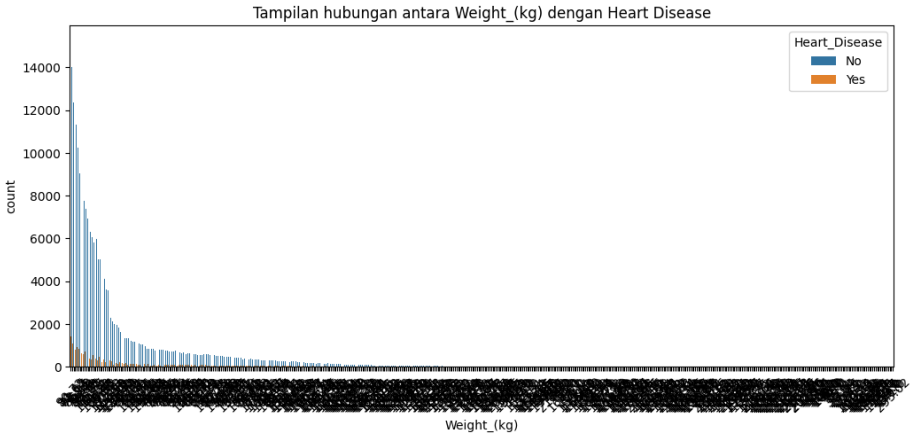
	Height_(cm)	Heart_Disease	Jumlah	Persentase
0	91.0	No	4	80.000000
1	91.0	Yes	1	20.000000
2	94.0	No	4	100.000000
3	96.0	No	1	100.000000
4	97.0	No	4	80.000000
5	97.0	Yes	1	20.000000
6	99.0	No	2	100.000000
7	100.0	No	2	100.000000
8	102.0	No	3	100.000000
9	103.0	No	1	100.000000

Gambar 5. 22 Jumlah atribut dalam Hubungan antara Variabel *Height (cm)* dengan Variabel *Heart Disease*

Berdasarkan Gambar 5.21 dan Gambar 5.22, dapat diketahui bahwa tinggi badan tertinggi variabel *Height (cm)* dan variabel *Heart Disease* yaitu setinggi 168 cm. Hubungan antara variabel *Height (cm)* dan variabel *Heart Disease* terdapat atribut *no* yang menandakan hubungan variabel tersebut tidak terdapat penyakit jantung. Kemudian juga terdapat atribut *yes* yang menandakan hubungan antara kedua variabel tersebut terdapat penyakit jantung. Jumlah masyarakat yang memiliki tinggi badan 168 cm yaitu sebanyak 27.119 orang dimana yang tidak terkena penyakit jantung sebanyak 25.144 orang dan yang terkena penyakit jantung sebanyak 1.975 orang. Tinggi badan 168 cm dapat terjadi risiko penyakit jantung yang apabila dipengaruhi oleh berbagai faktor seperti gaya hidup maupun faktor kesehatan yang lainnya.

5.1.1.12 Analisis Hubungan Variabel *Weight (kg)* dengan Variabel *Heart Disease*

Hubungan variabel *weight (kg)* dengan variabel *heart disease* yaitu untuk menjelaskan berat badan masyarakat yang menderita penyakit jantung. Atribut yang terdapat pada variabel ini mulai dari 24.95 kg, 25.40 kg, 26.31 kg, 26.76 kg, hingga 293.02 kg. Hal ini bertujuan untuk mengetahui berat badan masyarakat yang menderita penyakit jantung. Pada sub-bab ini disajikan dengan gambar 10 baris teratas, hal ini dikarenakan jumlah baris dalam variabel yang banyak dengan jumlah sebanyak 865 baris. Hubungan variabel *weight (kg)* dengan variabel *heart disease* dapat dilihat pada lampiran 3 pada halaman C-2. Berikut ini merupakan gambar 10 baris teratas yang menggambarkan hubungan variabel *weight (kg)* dengan variabel *heart disease*:



Gambar 5. 23 Tampilan Hubungan antara Variabel *Weight (kg)* dengan Variabel *Heart Disease*

Jumlah Variabel Weight_(kg) yang berhubungan dengan Variabel Heart Disease:

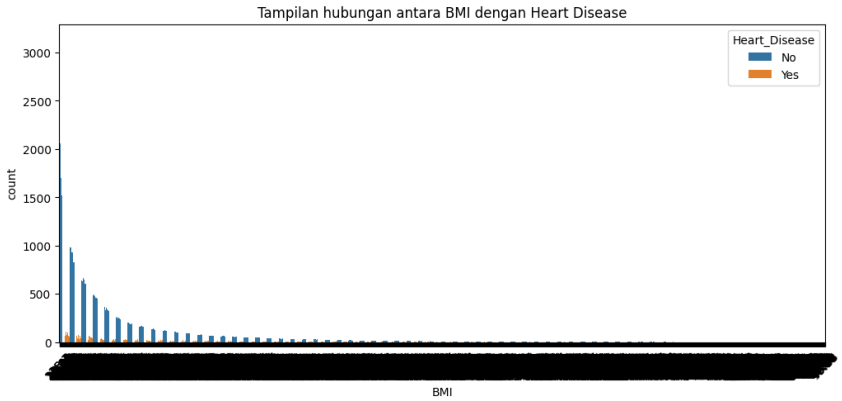
	Weight_(kg)	Heart_Disease	Jumlah	Persentase
0	24.95	No	1	100.000000
1	25.40	No	1	100.000000
2	26.31	No	1	100.000000
3	26.76	No	1	100.000000
4	27.22	No	1	100.000000
5	29.94	No	1	100.000000
6	30.00	No	1	50.000000
7	30.00	Yes	1	50.000000
8	30.84	Yes	1	100.000000
9	31.75	No	10	83.333333

Gambar 5. 24 Jumlah atribut dalam Hubungan antara Variabel *Weight (kg)* dengan Variabel *Heart Disease*

Berdasarkan Gambar 5.23 dan Gambar 5.24, dapat diketahui bahwa berat badang terberat antara variabel *Weight (kg)*dan variabel Heart Disease yaitu seberat 90.72 kg. dalam hubungan hubungan variabel *Weight (kg)* dengan variabel Heart Disease terdapat atribut *no* yang menandakan hubungan variabel tersebut tidak terdapat penyakit jantung. Kemudian juga terdapat atribut *yes* yang menandakan hubungan antara kedua variabel tersebut terdapat penyakit jantung. Jumlah orang yang memiliki berat badan seberat 90.72 kg sebanyak 16.614 orang yang terdiri dari yang tidak terkena penyakit jantung sebanyak 15.195 orang dengan persentase 91% dan yang terkena penyakit jantung sebanyak 1.419 orang dengan persentase 9%. Berat badan 90.72 kg dapat terjadi risiko kesehatan mengenai penyakit jantung yang dapat dipengaruhi melalui berbagai faktor seperti gaya hidup maupun faktor Kesehatan yang lainnya.

5.1.1.13 Analisis Hubungan Variabel *BMI* dengan Variabel *Heart Disease*

Hubungan variabel *BMI* dengan variabel *heart disease* yaitu untuk menjelaskan *BMI* (*Body Mass Index*) atau indeks masa tubuh dengan masyarakat yang menderita penyakit jantung. BMI digambarkan dengan berat badan dengan kondisi yang ideal. Atribut yang terdapat pada variabel ini mulai dari 12.02 Kg/m^2, 12.05 Kg/m^2, 12.11 Kg/m^2, 12.12 Kg/m^2 hingga 99.33 Kg/m^2. Hal ini bertujuan untuk mengetahui masa tubuh yang ideal masyarkat dengan penderita penyakit jantung. Pada sub-bab ini disajikan dengan gambar 10 baris teratas, hal ini dikarenakan jumlah baris dalam variabel yang banyak dengan jumlah sebanyak 5681 baris. Hubungan variabel *BMI* dengan variabel *heart disease* dapat dilihat pada lampiran 6 pada halaman C-3. Berikut ini merupakan gambar 10 baris teratas yang menggambarkan hubungan variabel *BMI* dengan variabel *heart disease*:



Gambar 5. 25 Tampilan Hubungan antara Variabel *BMI* dengan Variabel *Heart Disease*

Jumlah Variabel BMI yang berhubungan dengan Variabel Heart Disease:

	BMI	Heart_Disease	Jumlah	Persentase
0	12.02	No	1	100.000000
1	12.05	No	1	100.000000
2	12.11	Yes	1	100.000000
3	12.12	No	1	100.000000
4	12.16	No	4	100.000000
5	12.17	No	1	100.000000
6	12.20	No	1	100.000000
7	12.21	No	2	100.000000
8	12.40	No	1	100.000000
9	12.48	No	1	100.000000

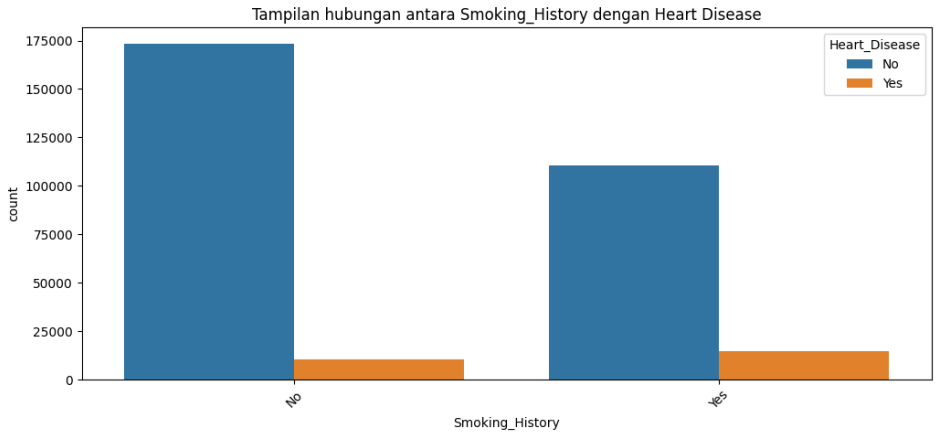
Gambar 5. 26 Jumlah atribut dalam Hubungan antara Variabel *BMI* dengan Variabel *Heart Disease*

Berdasarkan Gambar 5.25 dan Gambar 5.26, dapat diketahui bahwa nilai BMI tertinggi antara variabel *BMI* dan variabel *Heart Disease* yaitu sebesar 26.63 Kg/m^2. Dalam hubungan antara variabel *BMI* dengan variabel *Heart Disease* terdapat atribut *no* yang menandakan hubungan variabel tersebut tidak terdapat penyakit jantung. Kemudian juga

terdapat atribut *yes* yang menandakan hubungan antara kedua variabel tersebut terdapat penyakit jantung. Jumlah nilai dalam BMI sebanyak 3.340 orang yang dimana terdiri dari 26.63 Kg/m^2 yang tidak terkena penyakit jantung sebanyak 3.137 orang dengan proporsi persentase sebesar 94%. Kemudian nilai BMI yang terkena penyakit jantung memiliki nilai sebesar 203 orang dengan persentase sebesar 6%. Nilai BMI sebesar 26.63 Kg/m^2 menandakan bahwa kondisi badan tersebut dalam keadaan kelebihan berat badan yang berada diatas ambang badan ideal. Sehingga terdapat risiko kesehatan yang akan ditimbulkan, khususnya risiko terhadap penyakit jantung. Akan tetapi risiko kesehatan yang ditimbulkan tidak separah pada tingkatan telah mencapai obesitas.

5.1.1.14 Analisis Hubungan Variabel *Smoking History* dengan Variabel *Heart Disease*

Hubungan variabel *smoking history* dengan variabel *heart disease* yaitu untuk menjelaskan sejarah masyarakat yang mengkonsumsi rokok dengan masyarakat yang menderita penyakit jantung. Sejarah konsumsi rokok menggambarkan rekam jejak masyarakat yang mengkonsumsi rokok. Atribut yang terdapat dalam *dataset* yang digunakan terdiri dari *yes* artinya masyarakat mengkonsumsi rokok dan *no* artinya masyarakat tidak mengkonsumsi rokok. Hal ini bertujuan untuk mengetahui rekam jejak masyarakat yang mengkonsumsi rokok dengan penyakit jantung. Berikut ini merupakan gambar yang menggambarkan hubungan variabel *smoking history* dengan variabel *heart disease*:



Gambar 5. 27 Tampilan Hubungan antara Variabel *Smoking History* dengan Variabel *Heart Disease*

```
\Jumlah Variabel Smoking_History yang berhubungan dengan Variabel Heart Disease:
  Smoking_History Heart_Disease  Jumlah  Persentase
0              No            No  173203   94.342284
1              No           Yes   10387    5.657716
2             Yes            No  110680   88.357389
3             Yes           Yes   14584   11.642611
```

Gambar 5. 28 Jumlah atribut dalam Hubungan antara Variabel *Smoking History* dengan Variabel *Heart Disease*

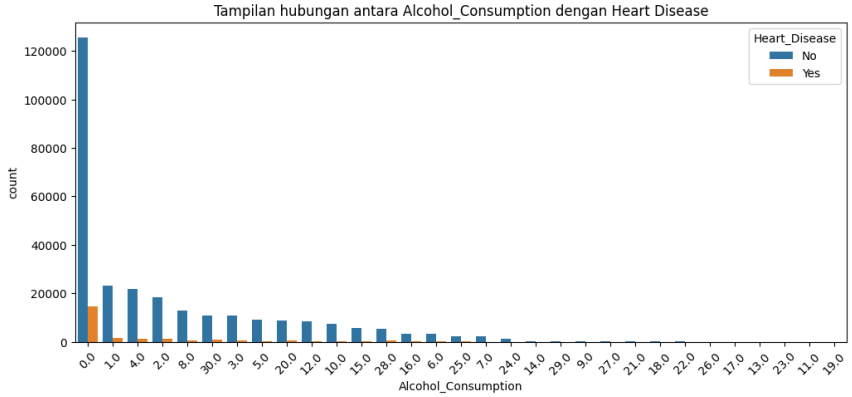
Berdasarkan Gambar 5.27 dan Gambar 5.28, dapat diketahui nilai tertinggi pada masing-masing atribut antara variabel diabetes dengan variabel heart disease yaitu *no* tidak terkena penyakit jantung dengan jumlah sebanyak 173.203 orang dengan proporsi persentase sebesar 94%. Kemudian pada atribut *yes* tidak terkena penyakit jantung dengan jumlah sebanyak 110.680 orang dengan proporsi persentase sebesar 88%. Kemudian jumlah masing-masing atribut dengan penyakit jantung yaitu pada atribut *no* dengan jumlah sebanyak 10.387 orang dengan proprose persentase sebesar 6%. Selanjutnya pada atribut *Yes* memiliki jumlah sebanyak 14.584 orang dengan proporsi persentase 12%.

Berdasarkan data tersebut dapat diketahui bahwa mayoritas masyarakat tidak merokok dan tidak menderita penyakit jantung. Proporsi persentase masyarakat yang tidak merokok dan tidak menderita penyakit jantung mencapai 94% dengan jumlah sebanyak 173.203 orang. Proporsi persentase dan jumlah yang berbeda dapat terjadi karena pola persebaran data yang berbeda. Kemudian jumlah masyarakat yang merokok dan memiliki proporsi persentase tertinggi yang menderita penyakit jantung yaitu terdapat pada atribut *yes* atau ya dalam merokok dan menderita penyakit jantung dengan jumlah sebanyak 14.584 orang dengan proporsi persentase sebesar 12%. Sehingga masyarakat yang merokok dan memiliki penyakit jantung menjadi perhatian khusus dalam penanganan penyakit jantung.

5.1.1.15 Analisis Hubungan Variabel *Alcohol Consumption* dengan Variabel *Heart Disease*

Hubungan variabel *alcohol consumption* dengan variabel *heart disease* yaitu untuk menjelaskan jumlah alkohol yang dikonsumsi oleh masyarakat dengan masyarakat yang menderita penyakit jantung. Atribut yang terdapat pada variabel ini terdiri dari frekuensi

mulai dari 0 kali, 1 kali, 2 kali, 3 kali, hingga 30 kali. Hal ini bertujuan untuk mengetahui jumlah konsumsi alkohol oleh masyarakat dengan penderita penyakit jantung. Pada sub-bab ini disajikan dengan gambar 10 baris teratas, hal ini dikarenakan jumlah baris dalam variabel yang banyak dengan jumlah sebanyak 60 baris. Hubungan variabel *alcohol consumption* dengan variabel *heart disease* dapat dilihat pada lampiran 8 pada halaman C-4. Berikut ini merupakan gambar 10 baris teratas yang menggambarkan hubungan variabel *alcohol consumption* dengan variabel *heart disease*:



Gambar 5. 29 Tampilan Hubungan antara Variabel *Alcohol Consumption* dengan Variabel *Heart Disease*

Jumlah Variabel Alcohol_Consumption yang berhubungan dengan Variabel Heart Disease:

	Alcohol_Consumption	Heart_Disease	Jumlah	Persentase
0	0.0	No	125443	89.434772
1	0.0	Yes	14819	10.565228
2	1.0	No	23185	92.803106
3	1.0	Yes	1798	7.196894
4	2.0	No	18531	93.875380
5	2.0	Yes	1209	6.124620
6	3.0	No	10831	94.354909
7	3.0	Yes	648	5.645091
8	4.0	No	21980	94.064279
9	4.0	Yes	1387	5.935721

Gambar 5. 30 Jumlah atribut dalam Hubungan antara Variabel *Alcohol Consumption* dengan Variabel *Heart Disease*

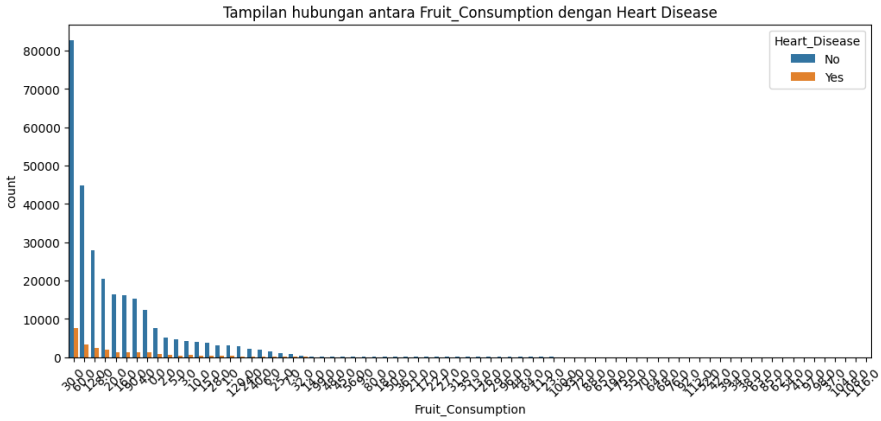
Berdasarkan Gambar 5.29 dan Gambar 5.30, dapat diketahui nilai tertinggi antara variabel Alcohol Consumption dengan variabel *Heart Disease* yaitu sebesar 0. Dalam hubungan antara Variabel *Alcohol Consumption* dengan variabel *Heart Disease* terdapat atribut *no* yang menandakan hubungan variabel tersebut tidak terdapat penyakit jantung. Kemudian juga terdapat atribut *yes* yang menandakan hubungan antara kedua variabel tersebut terdapat penyakit jantung.

Jumlah nilai dalam Variabel *Alcohol Consumption* dengan Variabel *Heart Disease* dengan nilai 0 yaitu sebanyak 140.262 orang yang teridiri dari yang tidak terkena penyakit jantung sebanyak 125.443 orang dengan proporsi persentase sebesar 89%, sedangkan yang terkena penyakit jantung sebanyak 14.819 orang dengan proporsi persentase sebesar 11%.

Nilai 0 dalam hubungan antara Variabel *Alcohol Consumption* dengan Variabel *Heart Disease* menandakan frekuensi alkohol dalam satu bulan sebanyak 0 kali atau tidak pernah mengkonsumsi alkohol. Konsumsi alkohol dapat mempengaruhi risiko kondisi kesehatan tubuh, khususnya terkait penyakit jantung. Akan tetapi walaupun tidak pernah mengkonsumsi alkohol masih dapat terjadi penyakit jantung yang dapat disebabkan oleh berbagai faktor seperti gaya hidup maupun faktor Kesehatan lainnya.

5.1.1.16 Analisis Hubungan Variabel *Fruit Consumption* dengan Variabel *Heart Disease*

Hubungan variabel *fruit consumption* dengan variabel *heart disease* yaitu untuk menjelaskan jumlah konsumsi buah oleh masyarakat dengan masyarakat yang menderita penyakit jantung. Atribut yang terdapat pada variabel ini terdiri dari frekuensi mulai dari 0 kali, 1 kali, 2 kali, 3 kali, hingga 120 kali. Hal ini bertujuan untuk mengetahui jumlah konsumsi buah oleh masyarakat dengan penderita penyakit jantung. Pada sub-bab ini disajikan dengan gambar 10 baris teratas, hal ini dikarenakan jumlah baris dalam variabel yang banyak dengan jumlah sebanyak 126 baris. Hubungan variabel *fruit consumption* dengan variabel *heart disease* dapat dilihat pada lampiran 9 pada halaman C-4. Berikut ini merupakan gambar 10 baris teratas yang menggambarkan hubungan variabel *fruit consumption* dengan variabel *heart disease*:



Gambar 5. 31 Tampilan Hubungan antara Variabel *Fruit Consumption* dengan Variabel *Heart Disease*

Jumlah Variabel Fruit_Consumption yang berhubungan dengan Variabel Heart Disease:

	Fruit_Consumption	Heart_Disease	Jumlah	Persentase
0	0.0	No	7507	90.087604
1	0.0	Yes	826	9.912396
2	1.0	No	2975	88.805970
3	1.0	Yes	375	11.194030
4	2.0	No	5192	89.640884
5	2.0	Yes	600	10.359116
6	3.0	No	4167	89.748008
7	3.0	Yes	476	10.251992
8	4.0	No	12409	90.662673
9	4.0	Yes	1278	9.337327

Gambar 5. 32 Jumlah atribut dalam Hubungan antara Variabel *Fruit Consumption* dengan Variabel *Heart Disease*

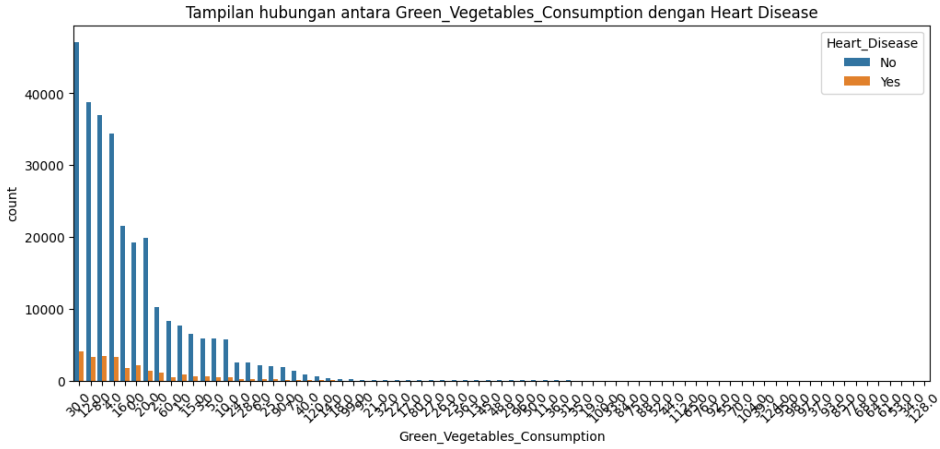
Berdasarkan Gambar 5.31 dan Gambar 5.32, dapat diketahui bahwa jumlah frekuensi tertinggi antara Variabel *Fruit Consumption* dengan Variabel *Heart Disease* yaitu sebesar 30. Dalam hubungan antara Variabel *Fruit Consumption* dengan Variabel *Heart Disease* terdapat atribut *no* yang menandakan hubungan variabel tersebut tidak terdapat penyakit jantung. Kemudian juga terdapat atribut *yes* yang menandakan hubungan antara kedua variabel tersebut terdapat penyakit jantung.

Jumlah nilai dalam Variabel *Fruit Consumption* dengan Variabel *Heart Disease* dengan nilai 30 yaitu sebanyak 90.273 orang yang teridiri dari yang tidak terkena penyakit jantung sebanyak 82.774 orang dengan proporsi persentase sebesar 92%, sedangkan yang terkena penyakit jantung sebanyak 7.499 orang dengan proporsi persentase sebesar 8%. Nilai 30 dalam hubungan antara Variabel *Fruit Consumption* dengan Variabel *Heart Disease* menandakan frekuensi sebanyak 30 kali. Frekuensi konsumsi buah sebanyak 30 kali menandakan bahwa mayoritas masyarakat paling tidak mengkonsumsi buah hampir

setiap dalam satu bulan. Akan tetapi walaupun mengkonsumsi buah masih dapat terjadi penyakit jantung yang dapat disebabkan oleh berbagai faktor.

5.1.1.17 Analisis Hubungan Variabel *Green Vegetables Consumption* dengan Variabel *Heart Disease*

Hubungan variabel *green vegetable consumption* dengan variabel *heart disease* yaitu untuk menjelaskan jumlah konsumsi sayuran hijau oleh masyarakat dengan masyarakat yang menderita penyakit jantung. Atribut yang terdapat pada variabel ini terdiri dari frekuensi mulai dari 0 kali, 1 kali, 2 kali, 3 kali, hingga 128 kali. Hal ini bertujuan untuk mengetahui jumlah konsumsi kentang goreng oleh masyarakat dengan penderita penyakit jantung. Pada sub-bab ini disajikan dengan gambar 10 baris teratas, hal ini dikarenakan jumlah baris dalam variabel yang banyak dengan jumlah sebanyak 110 baris. Hubungan variabel *green vegetable consumption* dengan variabel *heart disease* dapat dilihat pada gambar lampiran 10 pada halaman C-5. Berikut ini merupakan gambar 10 baris teratas yang menggambarkan hubungan variabel *green vegetable consumption* dengan variabel *heart disease*:



Gambar 5. 33 Tampilan Hubungan antara Variabel *Green Vegetables Consumption* dengan Variabel *Heart Disease*

Jumlah Variabel Green_Vegetables_Consumption yang berhubungan dengan Variabel Heart Disease:

	Green_Vegetables_Consumption	Heart_Disease	Jumlah	Persentase
0	0.0	No	19265	90.069662
1	0.0	Yes	2124	9.930338
2	1.0	No	7664	90.048173
3	1.0	Yes	847	9.951827
4	2.0	No	10293	90.687225
5	2.0	Yes	1057	9.312775
6	3.0	No	5899	91.301656
7	3.0	Yes	562	8.698344
8	4.0	No	34456	91.373412
9	4.0	Yes	3253	8.626588

Gambar 5. 34 Jumlah atribut dalam Hubungan antara Variabel *Green Vegetables Consumption* dengan Variabel *Heart Disease*

Berdasarkan Gambar 5.33 dan Gambar 5.34, dapat diketahui bahwa jumlah frekuensi tertinggi antara Variabel *Green Vegetables Consumption* dengan Variabel *Heart Disease* yaitu sebesar 30. Dalam hubungan antara Variabel *Green Vegetables Consumption* dengan Variabel *Heart Disease* terdapat atribut *no* yang menandakan hubungan variabel tersebut tidak terdapat penyakit jantung. Kemudian juga terdapat atribut *yes* yang menandakan hubungan antara kedua variabel tersebut terdapat penyakit jantung.

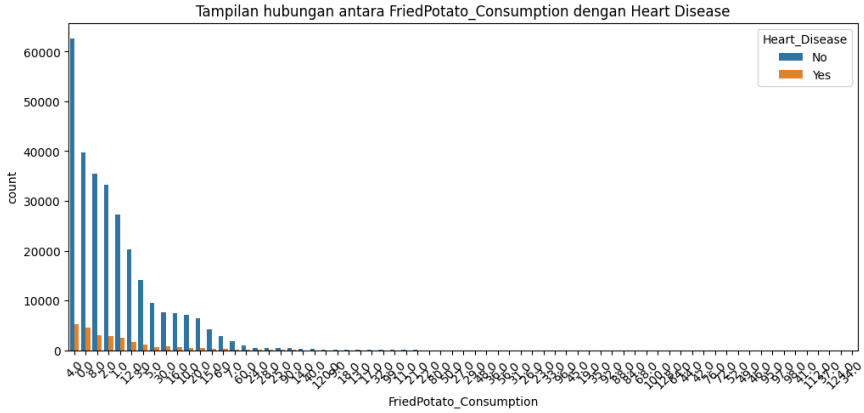
Jumlah nilai dalam Variabel *Green Vegetables Consumption* dengan Variabel *Heart Disease* dengan nilai 30 yaitu sebanyak 51.162 orang yang teridiri dari yang tidak terkena penyakit jantung sebanyak 47.122 orang dengan proporsi persentase sebesar 92%, sedangkan yang terkena penyakit jantung sebanyak 4.040 orang dengan proporsi persentase sebesar 8%.

Nilai 30 dalam hubungan antara Variabel *Green Vegetables Consumption* dengan Variabel *Heart Disease* menandakan frekuensi sebanyak 30 kali. Frekuensi konsumsi buah sebanyak 30 kali menandakan bahwa mayoritas masyarakat paling tidak mengkonsumsi buah hampir setiap dalam satu bulan. Akan tetapi walaupun mengkonsumsi buah masih dapat terjadi penyakit jantung yang dapat disebabkan oleh berbagai faktor.

5.1.1.18 Analisis Hubungan Variabel *Fried Potato Consumption* dengan Variabel *Heart Disease*

Hubungan variabel *fried potato consumption* dengan variabel *heart disease* yaitu untuk menjelaskan jumlah konsumsi kentang goreng oleh masyarakat dengan masyarakat yang menderita penyakit jantung. Atribut yang terdapat pada variabel ini terdiri dari frekuensi 0 kali, 1 kali, 2 kali, 3 kali, hingga 128 kali. Hal ini bertujuan untuk mengetahui jumlah

konsumsi kentang goreng oleh masyarakat dengan penderita penyakit jantung. Pada sub-bab ini disajikan dengan gambar 10 baris teratas, hal ini dikarenakan jumlah baris dalam variabel yang banyak dengan jumlah sebanyak 126 baris. Hubungan variabel *fruit consumption* dengan variabel *heart disease* dapat dilihat pada lampiran 11 pada halaman C-5. Berikut ini merupakan gambar 10 baris teratas yang menggambarkan hubungan variabel *fruit consumption* dengan variabel *heart disease*:



Gambar 5. 35 Tampilan Hubungan antara Variabel *Fried Potato Consumption* dengan Variabel *Heart Disease*

Jumlah Variabel FriedPotato_Consumption yang berhubungan dengan Variabel Heart Disease:

	FriedPotato_Consumption	Heart_Disease	Jumlah	Persentase
0	0.0	No	39723	89.803992
1	0.0	Yes	4510	10.196008
2	1.0	No	27325	91.461374
3	1.0	Yes	2551	8.538626
4	2.0	No	33289	92.226070
5	2.0	Yes	2806	7.773930
6	3.0	No	14190	92.461067
7	3.0	Yes	1157	7.538933
8	4.0	No	62627	92.325269
9	4.0	Yes	5206	7.674731

Gambar 5. 36 Jumlah atribut dalam Hubungan antara Variabel *Fried Potato Consumption* dengan Variabel *Heart Disease*

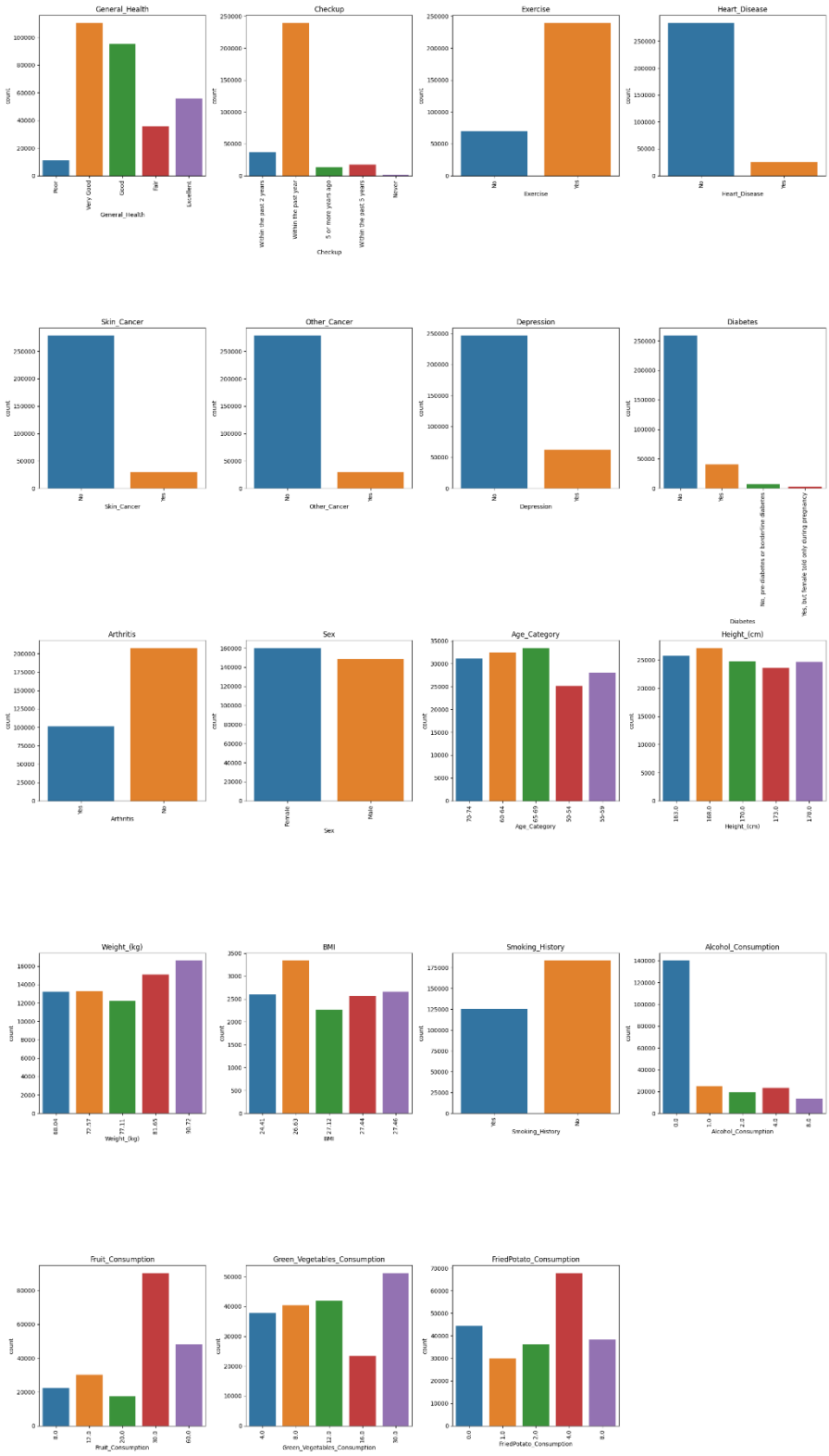
Berdasarkan Gambar 5.35 dan Gambar 5.36, dapat diketahui bahwa jumlah frekuensi tertinggi antara Variabel *Fried Potato Consumption* dengan Variabel *Heart Disease* yaitu sebesar 4. Dalam hubungan antara Variabel *Fried Potato Consumption* dengan Variabel *Heart Disease* terdapat atribut *no* yang menandakan hubungan variabel tersebut tidak terdapat penyakit jantung. Kemudian juga terdapat atribut *yes* yang menandakan hubungan antara kedua variabel tersebut terdapat penyakit jantung.

Jumlah nilai dalam Variabel *Fried Potato Consumption* dengan Variabel *Heart Disease* dengan nilai 4 yaitu sebanyak 67.833 orang yang teridiri dari yang tidak terkena

penyakit jantung sebanyak 62.627 orang dengan proporsi persentase sebesar 92%, sedangkan yang terkena penyakit jantung sebanyak 5.206 orang dengan proporsi persentase sebesar 8%. Nilai 4 kali menandakan bahwa frekuensi konsumsi kentang goreng dalam satu bulan sebanyak 4 kali, dimana paling tidak dalam 1 minggu konsumsi kentang goreng sebanyak 1 kali. Akan tetapi dengan konsumsi kentang goreng dengan jumlah yang rendah masih terdapat ancaman penyakit jantung yang disebabkan oleh berbagai faktor.

5.1.2 Analisis Hasil *Exploratory Data Analysis* (EDA)

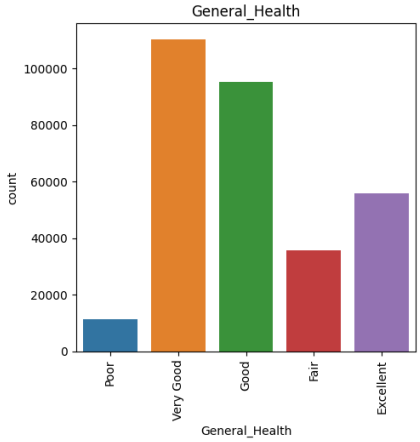
Exploratory Data Analysis (EDA) dalam penelitian ini bertujuan untuk melakukan eksplorasi data yang nantinya dilakukan visualisasi pada masing-masing variabel. Visualisasi yang dilakukan sebanyak 19 variabel dengan jenis diagram berupa *barchart* yang mudah dalam melakukan perbandingan maupun mengidentifikasi pola data, baik antar variabel maupun antar entititas. Berikut merupakan hasil visualisasi dengan menggunakan EDA:



Gambar 5. 37 Visualisasi *Exploratory Data Analysis*

Berdasarkan Gambar 5.2, terdapat 19 variabel yang divisualisaikan kedalam bentuk *barchart* memiliki nilai yang berbeda-beda yang ditampilkan dengan 5 nilai teratas dalam variabel. Berikut merupakan penjelasan masing-masing variabel:

5.1.2.1 Analisis EDA Variabel *General Health*



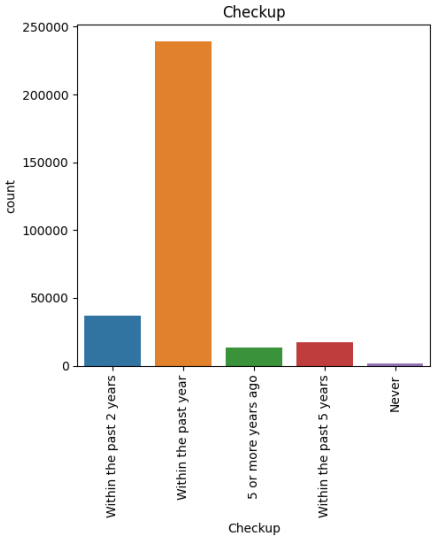
Gambar 5. 38 Analisis EDA Variabel *General Health*

```
Jumlah nilai pada variabel 'General_Health':
Very Good    110395
Good          95364
Excellent     55954
Fair          35810
Poor          11331
```

Gambar 5. 39 Jumlah atribut Variabel *General Health*

Berdasarkan gambar 5.38 dan Gambar 5.39 maka dapat diketahui jumlah atribut paling banyak dalam variabel general health yaitu *very good* sebanyak 110.395 orang, kemudian *good* sebanyak 95.364, *excellent* sebanyak 55.954 orang, *fair* sebanyak 35.810 orang, dan *poor* sebanyak 11.331 orang. Melalui hasil diatas maka dapat dikatakan bahwa secara mayoritas kondisi kesehatan masyarakat sudah baik. Hal ini dapat dilihat pada masyarakat dengan nilai atribut *very good*, *good*, dan *excellent* mencapai lebih dari setengah jumlah total masyarakat yaitu lebih dari 200.000 orang. Akan tetapi masih terdapat masyarakat dengan kondisi kesehatan yang kurang baik, khususnya pada atribut *fair* dan *poor* yang dapat menciptakan kesenjangan kondisi kesehatan di masyarakat. Sehingga dapat dikatakan pola data pada variabel *general health* menciptakan pola data dengan kesehatan yang relatif sehat.

5.1.2.2 Analisis EDA Variabel *Checkup*



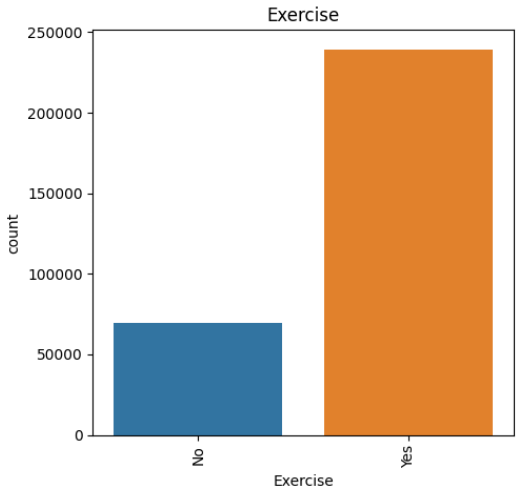
Gambar 5. 40 Analisis EDA Variabel *Checkup*

```
Jumlah nilai pada variabel 'Checkup':
Within the past year       239371
Within the past 2 years     37213
Within the past 5 years     17442
5 or more years ago         13421
Never                        1407
```

Gambar 5. 41 Jumlah atribut Variabel *Checkup*

Berdasarkan Gambar 5.40 dan Gambar 5.41 maka dapat diketahui jumlah atribut paling banyak dalam variabel *checkup* yaitu *within the past year* sebanyak 239.371 orang, kemudian *within the past 2 years* sebanyak 37.213 orang, *within the past 5 years* sebanyak 17.442 orang, *5 or more years ago* sebanyak 13.421 orang, dan *never* sebanyak 1.407 orang. Melalui hasil tersebut dapat diketahui bahwa mayoritas masyarakat melakukan pemerikasaan kesehatan secara rutin hampir tiap tahunnya. Hal ini dapat dilihat pada jumlah masyarakat yang melakukan pemeriksaan kesehatan rutin paling tidak selama 5 tahun sekali mencapai lebih dari 250.000 orang. Akan tetapi terdapat perbedaan yang cukup signifikan dengan orang yang tidak pernah melakukan pemeriksaan secara ruting yaitu sebanyak 1.407 orang.

5.1.2.3 Analisis EDA Variabel *Exercise*



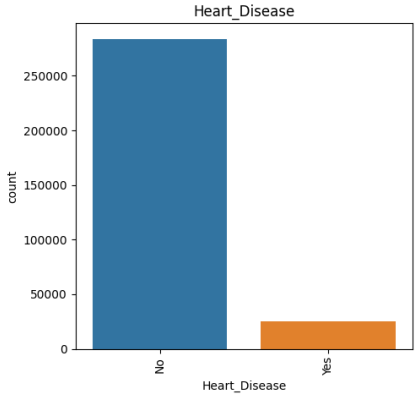
Gambar 5. 42 Analisis EDA Variabel *Exercise*

```
Jumlah nilai pada variabel 'Exercise':
Yes    239381
No      69473
```

Gambar 5. 43 Jumlah atribut Variabel *Exercise*

Berdasarkan Gambar 5.42 dan Gambar 5.43 maka dapat diketahui jumlah atribut paling banyak dalam variabel *exercise* yaitu *yes* sebanyak 239.381 orang dan *no* sebanyak 69.473 orang. Melalui hasil tersebut dapat dikatakan bahwa mayoritas masyarakat melakukan kegiatan aktivitas fisik yang jumlahnya mencapai lebih dari setengah total masyarakat yaitu lebih dari 230.000 orang. Akan tetapi masih terdapat sebagian kecil masyarakat yang tidak melakukan kegiatan aktivitas fisik yang jumlahnya hampir mencapai 70.000 orang dari total masyarakat secara keseluruhan.

5.1.2.4 Analisis EDA Variabel *Heart Disease*



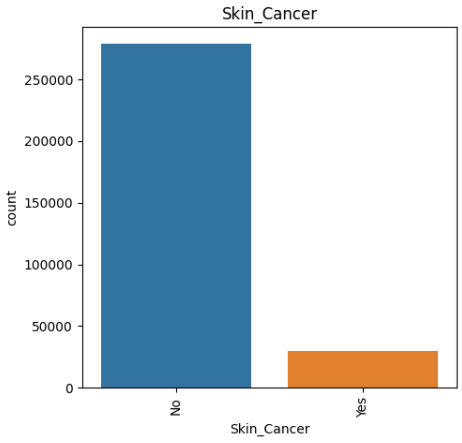
Gambar 5. 44 Analisis EDA Variabel *Heart Disease*

```
Jumlah nilai pada variabel 'Heart_Disease':
No     283883
Yes     24971
```

Gambar 5. 45 Jumlah atribut Variabel *Heart Disease*

Berdasarkan Gambar 5.44 dan Gambar 5.45 maka dapat diketahui jumlah atribut paling banyak dalam variabel *heart disease* yaitu *no* sebanyak 283.883 orang dan *yes* sebanyak 24.971 orang. Melalui hasil tersebut dapat dikatakan bahwa mayoritas masyarakat tidak mengalami penyakit jantung dengan jumlah lebih dari 250.000 orang. Sedangkan jumlah masyarakat yang mengalami penyakit jantung dengan jumlah hampir mencapai 25.000 orang dari total masyarakat. Sehingga dapat dikatakan mayoritas masyarakat bebas dari penyakit jantung.

5.1.2.5 Analisis EDA Variabel *Skin Cancer*



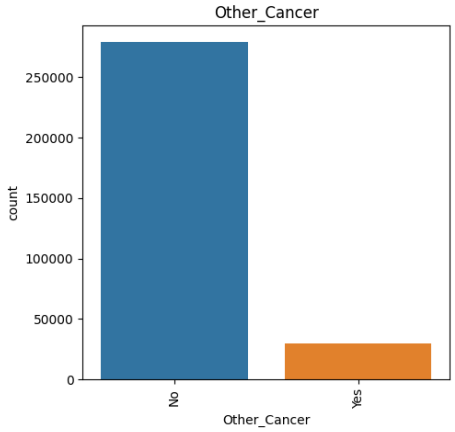
Gambar 5. 46 Analisis EDA Variabel *Skin Cancer*

```
Jumlah nilai pada variabel 'Skin_Cancer':
No     278860
Yes     29994
```

Gambar 5. 47 Jumlah atribut Variabel *Skin Cancer*

Berdasarkan Gambar 5.46 dan Gambar 5.47 maka dapat diketahui jumlah atribut paling banyak dalam variabel *skin cancer* yaitu *no* sebanyak 278.860 orang dan *yes* sebanyak 29.994 orang. Melalui hasil tersebut dapat dikatakan bahwa mayoritas masyarakat tidak menderita penyakit kanker kulit denganjumlah lebih dari 270.000 orang. Sedangkan jumlah masyarakat yang menderita penyakit kanker kulit dengan jumlah hampir mencapai 30.000 orang dari total masyarakat. Sehingga dapat dikatakan bahwa sebagian besar masyarakat tidak menderita penyakit kanker kulit.

5.1.2.6 Analisis EDA Variabel *Other Cancer*



Gambar 5. 48 Analisis EDA Variabel *Other Cancer*

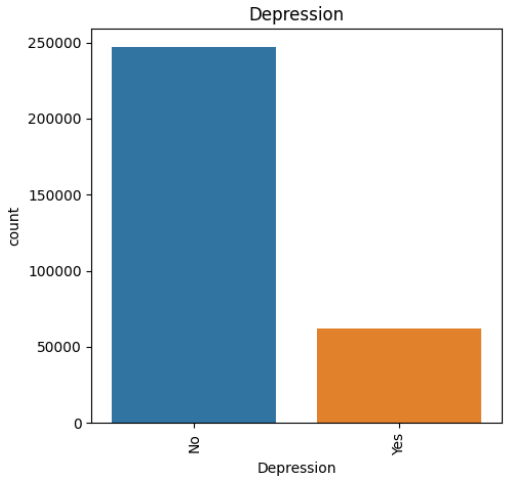
```
Jumlah nilai pada variabel 'Other_Cancer':
No     278976
Yes     29878
```

Gambar 5. 49 Jumlah atribut Variabel *Other Cancer*

Berdasarkan Gambar 5.48 dan Gambar 5.49 maka dapat diketahui jumlah atribut paling banyak dalam variabel *other cancer* yaitu *no* sebanyak 278.976 orang dan *yes* sebanyak 29.878 orang. Melalui hasil tersebut dapat dikatakan bahwa mayoritas masyarakat tidak menderita penyakit kanker dengan jenis yang lain selain kanker kulit dengan jumlah lebih dari 270.000 orang. Sedangkan jumlah masyarakat yang menderita penyakit kanker dengan jenis yang lain selain kanker kulit dengan jumlah hampir mencapai 30.000 orang

dari total masyarakat. Sehingga dapat dikatakan bahwa sebagian besar masyarakat tidak menderita penyakit kanker dengan jenis yang lainnya.

5.1.2.7 Analisis EDA Variabel *Depression*



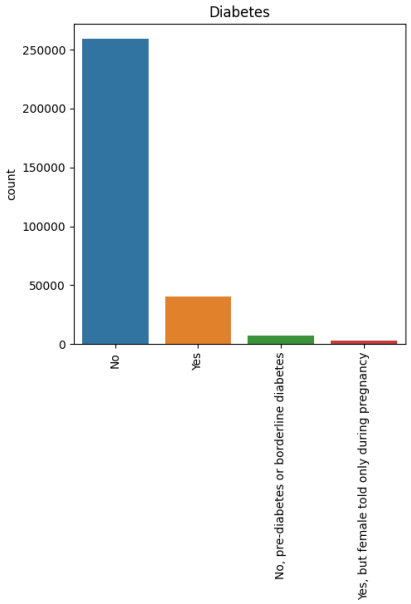
Gambar 5. 50 Analisis EDA Variabel *Depression*

```
Jumlah nilai pada variabel 'Depression':
No     246953
Yes     61901
```

Gambar 5. 51 Jumlah atribut Variabel *Depression*

Berdasarkan Gambar 5.50 dan Gambar 5.51 maka dapat diketahui jumlah atribut paling banyak dalam variabel *depression* yaitu *no* sebanyak 246.953 orang dan *yes* sebanyak 61.901 orang. Melalui hasil tersebut dapat dikatakan bahwa mayoritas masyarakat tidak mengalami penyakit depresi dengan jumlah hampir 250.000 orang dari total masyarakat. Akan tetaou jumlah masyarakat yang mengalami penyakit depresi mencapai jumlah lebih dari 60.000 orang dari total masyarakat. Maka dapat dikatakan bahwa sebagian masyarakat tidak mengalami penyakit depresi.

5.1.2.8 Analisis EDA Variabel *Diabetes*



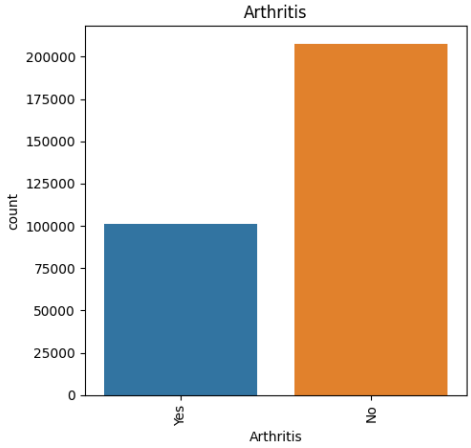
Gambar 5. 52 Analisis EDA Variabel *Diabetes*

```
Jumlah nilai pada variabel 'Diabetes':
No                                       259141
Yes                                       40171
No, pre-diabetes or borderline diabetes    6896
Yes, but female told only during pregnancy 2646
```

Gambar 5. 53 Jumlah atribut Variabel *Diabetes*

Berdasarkan Gambar 5.52 dan Gambar 5.53 maka dapat diketahui jumlah atribut paling banyak dalam variabel *arthritis* yaitu *no* sebanyak 207.783 orang, kemudian *yes* sebanyak 101.071 orang, *no, pre-diabetes or borderline diabetes* sebanyak 6.896 orang, dan *yes, but female told only during pregnancy* sebanyak 2.646 orang. Melalui hasil tersebut dapat dikatakan mayoritas masyarakat tidak mengalami penyakit diabetes dengan jumlah mencapai lebih dari 250.000 orang dari total masyarakat. Akan tetapi masih terdapat masyarakat yang mengalami penyakit diabetes dengan jumlah lebih dari 40.000 orang dari total masyarkat. Sehingga dapat dikatakan masyarakat bahwa sebagian masyarakat terbebas dari penyakit diabetes.

5.1.2.9 Analisis EDA Variabel *Arthritis*



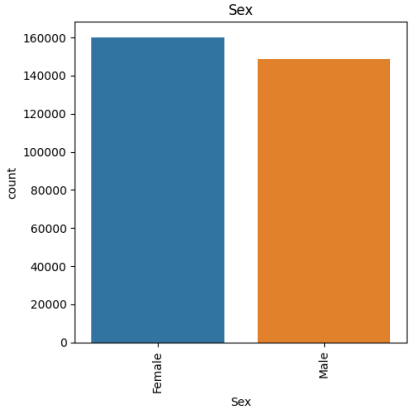
Gambar 5. 54 Analisis EDA Variabel *Arthritis*

```
Jumlah nilai pada variabel 'Arthritis':
No     207783
Yes    101071
Name: Arthritis, dtype: int64
```

Gambar 5. 55 Jumlah atribut Variabel *Arthritis*

Berdasarkan Gambar 5.54 dan Gambar 5.55, maka dapat diketahui jumlah atribut paling banyak dalam variabel *arthritis* yaitu *no* sebanyak 207.783 orang dan *yes* sebanyak 101.071 orang. Melalui hasil tersebut dapat dikatakan bahwa sebagian besar masyarakat menderita penyakit arthritis dengan jumlah lebih dari 200.000 orang. Sedangkan masih terdapat masyarkat yang menderita penyakit arthritis dengan jumlah lebih dari 100.000 orang. Sehingga dapat dikatakan bahwa mayoritas masyarakat tidak menderita penyakit arthritis.

5.1.2.10 Analisis EDA Analisis EDA Variabel *Sex*



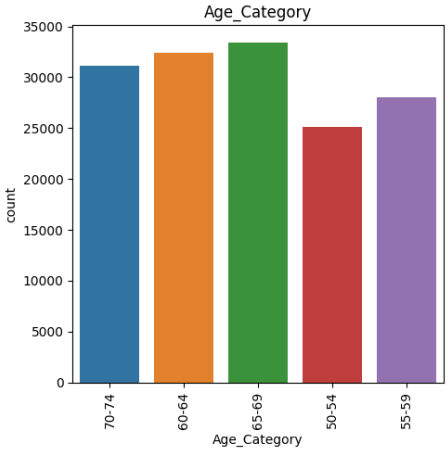
Gambar 5. 56 Analisis EDA Variabel *Sex*

```
Jumlah nilai pada variabel 'Sex':
Female    160196
Male      148658
```

Gambar 5. 57 Jumlah atribut Variabel *Sex*

Berdasarkan Gambar 5.56 dan Gambar 5.57 maka dapat diketahui jumlah atribut paling banyak dalam variabel *sex* yaitu *female* sebanyak 160.196 orang dan *male* sebanyak 148.658 orang. Berdasarkan hasil tersebut dapat diketahui jumlah masyarakat dengan jenis kelamin perempuan maupun laki-laki dengan jumlah yang tidak terlalu jauh dimana jenis kelamin perempuang memiliki jumlah yang lebih banyak dibandingkan dengan jenis kelamin laki-laki.

5.1.2.11 Analisis EDA Variabel *Age Category*



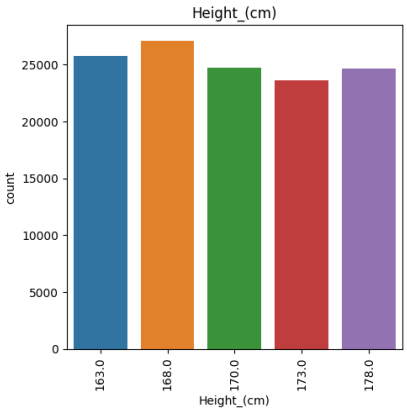
Gambar 5. 58 Analisis EDA Variabel *Age Category*

```
Jumlah nilai pada variabel 'Age_Category':
65-69    33434
60-64    32418
70-74    31103
55-59    28054
50-54    25097
80+      22271
40-44    21595
45-49    20968
75-79    20705
35-39    20606
18-24    18681
30-34    18428
25-29    15494
```

Gambar 5. 59 Jumlah atribut Variabel *Age Category*

Berdasarkan Gambar 5.58 dan Gambar 5.59 maka dapat diketahui jumlah atribut paling banyak dalam variabel *age category* yaitu 18-24 tahun sebanyak 18.681 orang, 25-29 tahun sebanyak 15.494 orang, 30-34 tahun sebanyak 18.428 orang, 35-39 tahun sebanyak 20.606 orang, 40-44 tahun sebanyak 21.595 orang, 45-49 tahun sebanyak 20.968 orang, 50-54 tahun sebanyak 25.097 orang, 55-59 tahun sebanyak 28.054 orang, 60-64 tahun sebanyak 32.418 orang, 65-69 tahun sebanyak 33.434 orang, 70-74 tahun 31.103 orang, 75-79 tahun sebanyak 20.705 orang, dan usia 80 tahun ke atas sebanyak 22.271 orang. Berdasarkan hasil diatas maka dapat diketahui bahwa kategori usia dengan jumlah terbanyak terdapat pada rentang usia 65-69 tahun. Sehingga dapat diketahui bahwa banyak kategori usia masyarakat yang tergolong sebagai lanjut usia yaitu terdapat pada rentang usia 65-69 tahum, 60-64 tahun, dan 70-74 tahun.

5.1.2.12 Analisis EDA Variabel *Height (cm)*



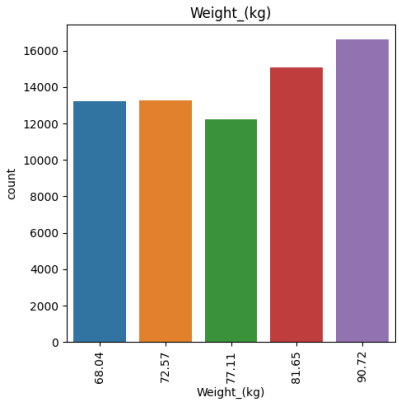
Gambar 5. 60 Analisis EDA Variabel *Height (cm)*

```
Jumlah nilai pada variabel 'Height_(cm)':
168.0    27119
163.0    25804
170.0    24739
178.0    24673
173.0    23591
165.0    23551
175.0    22059
183.0    22008
160.0    20829
157.0    19377
```

Gambar 5. 61 Jumlah atribut Variabel *Height (cm)*

Berdasarkan Gambar 5.60 dan Gambar 5.61 maka dapat diketahui jumlah atribut paling banyak dalam variabel *height (cm)* dengan posisi 5 teratas yaitu 168 cm sebanyak 27.119 orang, 163 cm sebanyak 25.804 orang, 170 cm sebanyak 24.739 orang, 178 cm sebanyak 24.673 orang, dan 173 cm sebanyak 23.591 orang. Jumlah keseluruhan pada variabel *height (cm)* mencapai 99 baris. Variabel *height (cm)* dapat dilihat pada lampiran 12 pada halaman C-6. Melalui hasil diatas dapat diketahui bahwa banyak masyarakat yang memiliki tinggi badan diatas 160 cm dengan jumlah lebih dari 100.00 orang. Sehingga dapat diketahui bagwa pola data variabel *height (cm)* membentuk pola data dengan skala tinggi badan masyarakat yang tinggi.

5.1.2.13 Analisis EDA Variabel *Weight (kg)*



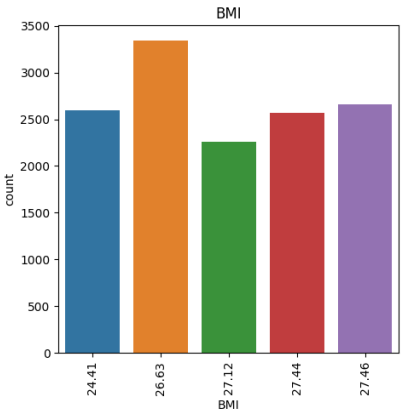
Gambar 5. 62 Analisis EDA Variabel *Weight (kg)*

```
Jumlah nilai pada variabel 'Weight_(kg)':
90.72     16614
81.65     15104
72.57     13263
68.04     13240
77.11     12216
86.18     11083
63.50      9583
79.38      9066
74.84      8559
99.79      8454
```

Gambar 5. 63 Jumlah atribut Variabel *Weight (kg)*

Berdasarkan Gambar 5.62 dan Gambar 5.63 maka dapat diketahui jumlah atribut paling banyak dalam variabel *weight (kg)* dengan posisi 5 teratas yaitu 90.72 kg sebanyak 16.614 orang, 81.65 kg sebanyak 15.104 orang, 72.57 kg sebanyak 13.263 orang, 68.04 kg sebanyak 13.240 orang, dan 77.11 kg sebanyak 12.216 orang. Variabel *weight (kg)* dapat dilihat pada lampiran 13 pada halaman C-6. Melalui hasil diatas dapat diketahui bahwa banyak masyarakat yang memiliki tinggi badan diatas 75 kg dengan jumlah lebih dari 100.00 orang. Sehingga dapat diketahui bagwa pola data variabel *weight (kg)* membentuk pola data dengan skala berat badan masyarakat yang cukup gemuk.

5.1.2.14 Analisis EDA Variabel *BMI*



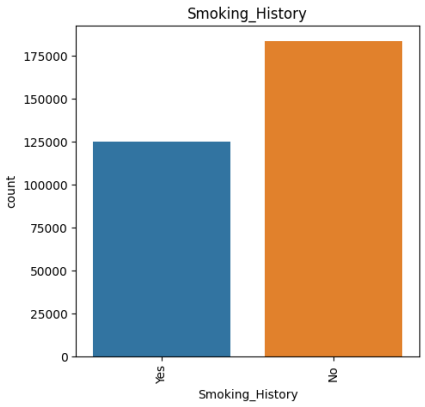
Gambar 5. 64 Analisis EDA Variabel *BMI*

```
Jumlah nilai pada variabel 'BMI':
26.63    3340
27.46    2658
24.41    2596
27.44    2567
27.12    2259
25.10    2068
32.28    1935
28.70    1894
29.53    1881
29.29    1853
```

Gambar 5. 65 Jumlah atribut Variabel *BMI*

Berdasarkan Gambar 5.64 dan Gambar 5.65 maka dapat diketahui jumlah atribut paling banyak dalam variabel *BMI* dengan posisi 5 teratas yaitu 26.63 Kg/m^2 sebanyak 3.340 orang, 27.46 Kg/m^2 sebanyak 2.658 orang, 24.41 Kg/m^2 sebanyak 2.596 orang, 27.44 Kg/m^2 sebanyak 2.567 orang, dan 27.12 Kg/m^2 sebanyak 2.259 orang. Jumlah variabel *BMI* dapat dilihat pada lampiran 14 pada halaman C-6. Melalui hasil diatas dapat diketahui bahwa banyak masyarakat yang memiliki tinggi badan diatas 75 kg dengan jumlah lebih dari 100.00 orang. Sehingga dapat diketahui bagwa pola data variabel *BMI* membentuk pola data dengan skala berat badan masyarakat yang cukup gemuk

5.1.2.15 Analis EDA Variabel *Smoking History*



Gambar 5. 66 Analisis EDA Variabel *Smoking History*

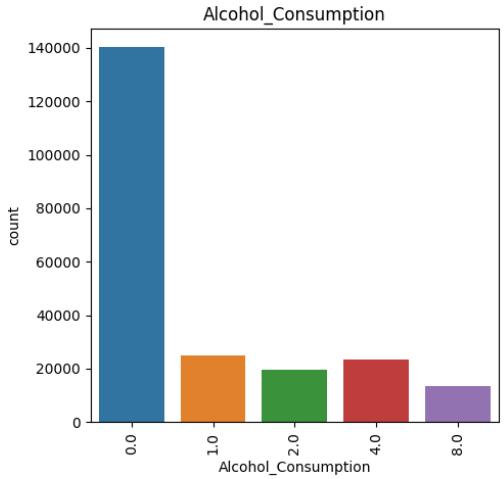
```
Jumlah nilai pada variabel 'Smoking_History':
No     183590
Yes    125264
```

Gambar 5. 67 Jumlah atribut Variabel *Smoking History*

Berdasarkan Gambar 5.66 dan Gambar 5.67 maka dapat diketahui jumlah atribut paling banyak dalam variabel *smoking history* yaitu *no* sebanyak 183.590 orang dan *yes* sebanyak 125.264 orang. Berdasarkan hasil tersebut dapat dikatakan bahwa masyarakat yang tidak merokok memiliki jumlah yang lebih banyak dibandingkan dengan

masyarakat yang merokok dengan jumlah selisih perbedaan mendapai lebih dari 50.000 orang. Sehingga dapat dikatakan bahwa banyak masyarakat yang tidak merokok.

5.1.2.16 Analisis EDA Variabel *Alcohol Consumption*



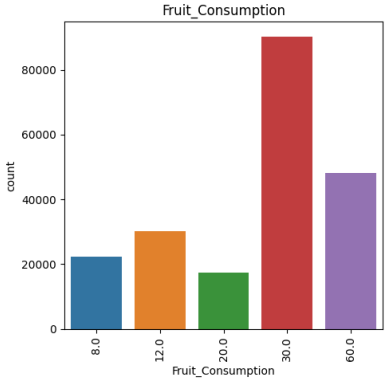
Gambar 5. 68 Analisis EDA Variabel *Alcohol Consumption*

```
Jumlah nilai pada variabel 'Alcohol_Consumption':
0.0     140262
1.0      24983
4.0      23367
2.0      19740
8.0      13438
30.0     11976
3.0      11479
5.0       9622
20.0      9372
12.0      8825
```

Gambar 5. 69 Jumlah atribut Variabel *Alcohol Consumption*

Berdasarkan gambar 5.68 dan Gambar 5.69 maka dapat diketahui jumlah atribut paling banyak dalam variabel *alcohol consumption* dengan posisi 5 teratas yaitu dengan frekuensi 0 kali sebanyak 140.262 orang, 1 kali sebanyak 24.983 orang, 4 kali sebanyak 23.367 orang, 2 kali sebanygak 19.740 orang, dan 8 kali sebanyak 13.438 orang. Jumlah variabel *alcohol consumption* dapaat dilihat pada lampiran 15 pada halaman C-7. Melalui hasil diatas dapat diketahui bahwa mayoritas masyarakat cenderung tidak mengonsumsi alkohol dengan nilai frekuensi 0 kali yang berjumlah sebanyak lebih dari 140.000 orang. Sehingga dapat diketahui bahwa pola data variabel *alcohol consumption* membentuk pola data dengan masyarakat mengonsumsi alkohol dengan jumlah yang sedikit.

5.1.2.17 Analisis EDA Variabel *Fruit Consumption*



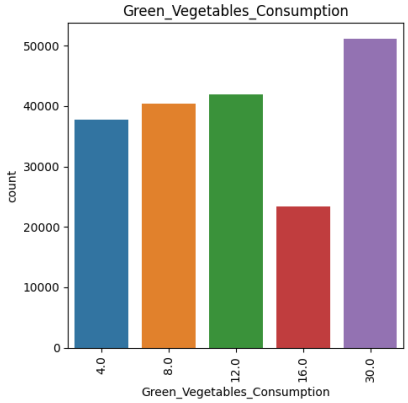
Gambar 5. 70 Analisis EDA Variabel *Fruit Consumption*

```
Jumlah nilai pada variabel 'Fruit_Consumption':
30.0     90273
60.0     48127
12.0     30259
8.0      22279
20.0     17476
16.0     17444
90.0     16567
4.0      13687
0.0       8333
2.0       5792
```

Gambar 5. 71 Jumlah atribut Variabel *Fruit Consumption*

Berdasarkan Gambar 5.70 dan Gambar 5.71 maka dapat diketahui jumlah atribut paling banyak dalam variabel *fruit consumption* dengan posisi 5 teratas yaitu dengan frekuensi 30 kali sebanyak 90.723 orang, 60 kali sebanyak 48.127 orang, 12 kali sebanyak 30.259 orang, 8 kali sebanyak 22.279 orang, dan 20 kali sebanyak 17.476 orang. Jumlah variabel *fruit consumption* dapat dilihat pada lampiran 16 pada halaman C-7. Melalui hasil diatas dapat diketahui bahwa mayoritas masyarakat mengonsumsi buah dengan frekuensi sebanyak 30 kali dengan jumlah lebih dari 90.000 orang. Sehingga dapat diketahui bahwa pola data variabel *fruit consumption* membentuk pola data dengan sebagian besar masyarakat mengonsumsi buah.

5.1.2.18 Analisis EDA Variabel *Green Vegetables Consumption*



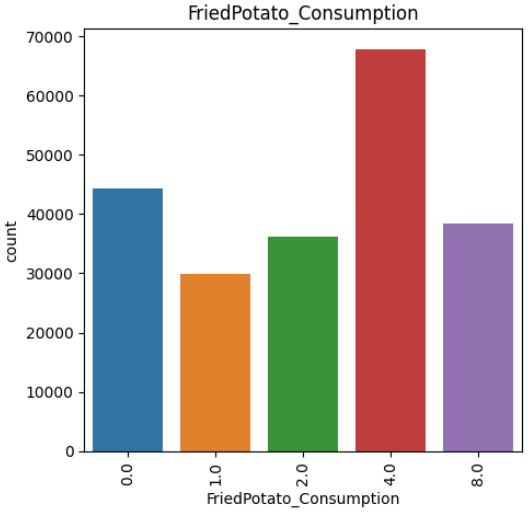
Gambar 5. 72 Analisis EDA Variabel *Green Vegetables Consumption*

```
Jumlah nilai pada variabel 'Green_Vegetables_Consumption':
30.0     51162
12.0     41979
8.0      40329
4.0      37709
16.0     23338
0.0      21389
20.0     21288
2.0      11350
60.0      8796
1.0       8511
```

Gambar 5. 73 Jumlah atribut Variabel *Green Vegetables Consumption*

Berdasarkan Gambar 5.72 dan Gambar 5.73 maka dapat diketahui jumlah atribut paling banyak dalam variabel *green vegetables consumption* dengan posisi 5 teratas yaitu dengan frekuensi 30 kali sebanyak 51.152 orang, 12 kali sebanyak 41.979 orang, 8 kali sebanyak 40.329 orang, 4 kali sebanyak 37.709 orang, dan 16 kali sebanyak 23.338 orang. Jumlah variabel *green vegetables consumption* dapat dilihat pada lampiran 17 pada halaman C-7. Melalui hasil diatas dapat diketahui bahwa mayoritas masyarakat mengonsumsi buat sebanyak 30 porsi dengan jumlah lebih dari 50.000 orang. Sehingga dapat diketahui bahwa pola data variabel *green vegetables consumption* membentuk pola data dengan sebagian besar masyarakat mengonsumsi sayuran hijau.

5.1.2.19 Analisis EDA Variabel *Fried Potato Consumption*



Gambar 5. 74 Analisis EDA Variabel *Fried Potato Consumption*

```
Jumlah nilai pada variabel 'FriedPotato_Consumption':
4.0     67833
0.0     44233
8.0     38366
2.0     36095
1.0     29876
12.0    21940
3.0     15347
5.0     10121
30.0     8434
16.0     8038
```

Gambar 5. 75 Jumlah atribut Variabel *Fried Potato Consumption*

Berdasarkan Gambar 5.74 dan Gambar 5.75, maka dapat diketahui jumlah atribut paling banyak dalam variabel *fried potato consumption* dengan posisi 5 teratas yaitu dengan frekuensi 4 kali sebanyak 67.833 orang, 0 kali sebanyak 44.233 orang, 8 kali sebanyak 38.366 orang, 2 kali sebanyak 36.095 orang, dan 1 kali sebanyak 29.876 orang. Jumlah variabel *fried potato consumption* dapat dilihat pada lampiran 18 pada halaman C-8. Melalui hasil diatas dapat diketahui bahwa mayoritas masyarakat mengonsumsi buat sebanyak 4 porsi dengan jumlah lebih dari 60.000 orang. Sehingga dapat diketahui bahwa pola data variabel *fried potato consumption* membentuk pola data dengan sebagian besar masyarakat mengonsumsi kentang goreng dengan jumlah yang rendah.

5.1.3 Analisis Hasil *Encoding Data*

Encoding data dalam penelitian ini dilakukan untuk mengubah nilai kategorikal dalam suatu variabel menjadi nilai numerik. Encoding data yang digunakan berupa encoding dengan jenis label encoder yang memberikan gambaran mengenai nilai numerik terdapat pada dataset diatas yang sebelumnya berbentuk variabel kategorikal. Variabel yang nilainya diubah yaitu dengan nilai non-numerik menjadi nilai numerik. Variabel yang nilainya diubah yaitu *age category, diabetes, general health, checkup, sex, skin cancer, other cancer, depression, arthritis, exercise,* dan *smoking history*.

Pada variabel *age category* nilai dalam atribut yang diubah yaitu 18-24 tahun menjadi 0, 25-29 tahun menjadi 1, 30-34 tahun menjadi 2, 35-39 tahun menjadi 3, 40-44 tahun menjadi 4, 45-49 tahun menjadi 5, 50-54 tahun menjadi 6, 55-59 tahun menjadi 7, 60-64 tahun menjadi 8, 65-69 tahun menjadi 9, 70-74 tahun menjadi 10, 75-79 tahun menjadi 11, dan usia 80 tahun keatas menjadi 12. Kemudian pada variabel *diabetes* nilai dalam atribut yang diubah yaitu *no* menjadi 0, *no, pre-diabetes or borderline diabetes* menjadi 1, *yes* menjadi 2, dan *yes, but female told only during pregnancy* menjadi 3.

Selanjutnya pada variabel *general health* nilai dalam atribut yang diubah yaitu *excellent* menjadi 0, *fair* menjadi 1, *good* menjadi 2, *poor* menjadi 3, dan *very good* menjadi 4. Pada variabel *checkup* nilai dalam atribut yang diubah yaitu *5 or more years ago* menjadi 0, *never* menjadi 1, *within the past 2 years* menjadi 2, *within the past 5 years* menjadi 3, dan *within the past year* menjadi 4.

Lalu pada variabel *sex* nilai dalam atribut yang diubah yaitu *female* menjadi 0 dan *male* menjadi 1. Pada variabel, *skin cancer, other cancer, depression, arthritis, exercise,* dan *smoking history* memiliki nilai atribut yang sama dengan nilai atribut yang diubah yaitu *no* menjadi 0 dan *yes* menjadi 1. Hasil *encoding data* dapat dilihat pada lampiran 20 dan 21 pada halaman C-8.

5.1.4 Analisis Hasil *Data Cleaning*

Data cleaning dilakukan untuk memeriksa *dataset* yang siap digunakan untuk melakukan pemodelan dengan *AutoGluon*. Data cleaning yang dilakukan dalam penelitian ini yaitu dengan melakukan *handling missing values, handling duplicate*, dan *feature scaling*.

5.1.4.1 Analisis Hasil *Handling Missing Values*

Handling missing values dilakukan untuk menangani nilai-nilai yang kosong pada dataset dengan memeriksa adanya nilai yang kosong. Pemeriksaan nilai yang kosong dilakukan pada masing-masing kolom. Berdasarkan pemeriksaan yang dilakukan tidak terdapat variabel yang memiliki nilai yang kosong. Hasil pemeriksaan *handling missing values* dapat dilihat pada lampiran 22 dan 23 pada halaman C-9.

5.1.4.2 Analisis Hasil *Handling Data Dupplicate*

Data duplicate dilakukan untuk memeriksa adanya data yang mengalami duplikasi dengan nilai yang sama pada setiap barisnya. Berdasarkan hasil pemeriksaan yang dilakukan maka dapat diketahui jumlah baris data yang mengalami duplikasi sebanyak 80 baris. Sehingga baris yang mengalami duplikasi harus dihapus agar tidak mempengaruhi hasil akhir pemodelan. Maka jumlah baris secara keseluruhan terdapat 308.854 baris berkurang menjadi 308.774. Hasil pemeriksaan baris yang mengalami duplikasi dapat dilihat pada lampiran 24 pada halaman C-9.

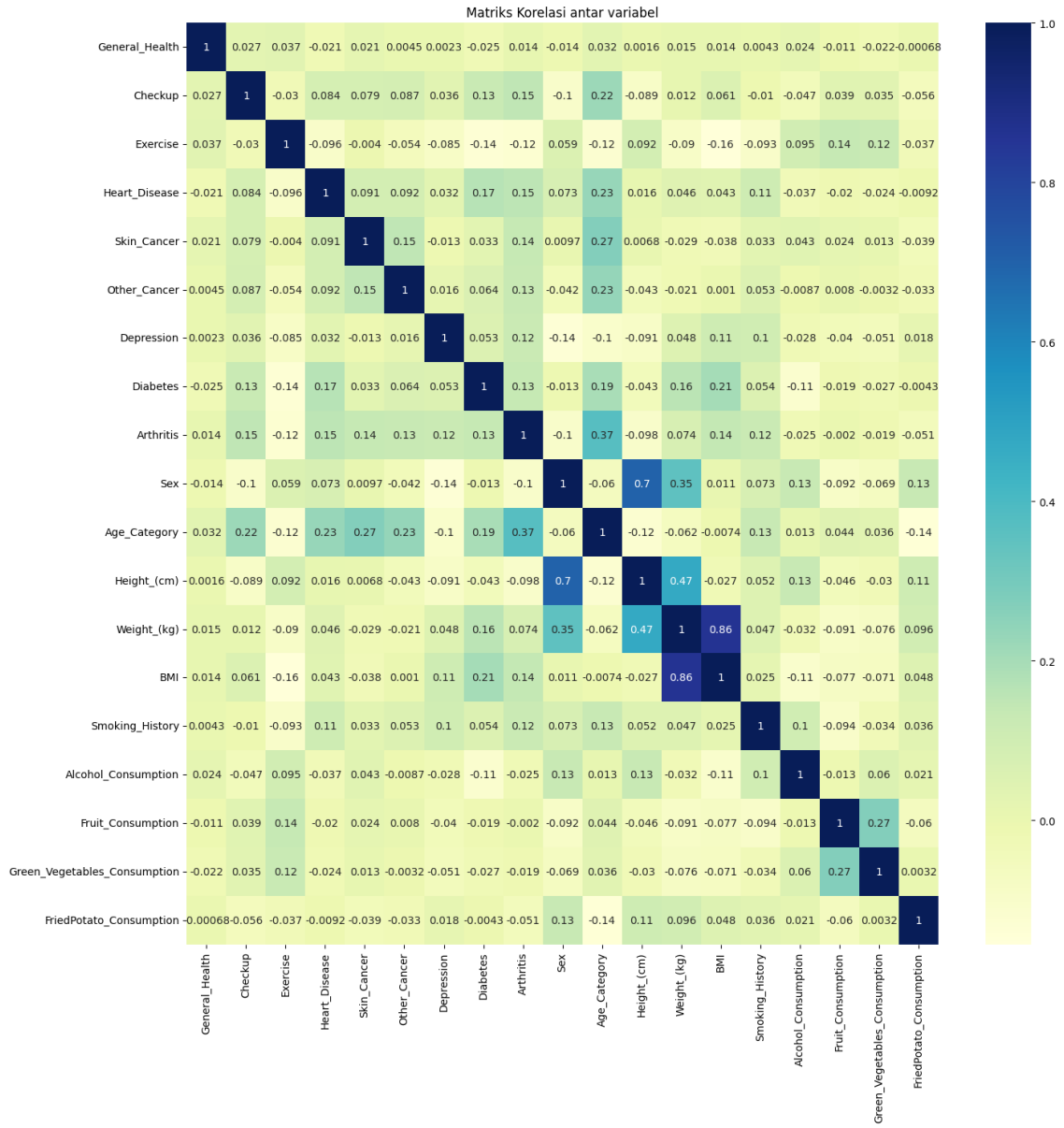
5.1.4.3 Analisis Hasil *Feature Scaling*

Feature scaling dilakukan untuk mengubah nilai variabel pada dataset menjadi skala dengan nilai konsisten yang dilakukan dengan metode *standard scaler*. Pemilihan metode *standard scaler* karena metode ini dapat menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan *min-max scaler*. *Min-Max Scaler* merupakan metode normalisasi yang dilakukan dengan menyesuaikan nilai fitur yang memiliki interval dari 0 hingga 1 dengan mempertahankan perbedaan yang terdapat dalam interval pada nilai (Ardana, 2023). Penelitian yang dilakukan oleh (Zoran Stojanoski, 2023) dengan judul *Comparative Analysis of Machine Learning Models for Diabetes Prediction* dengan hasil pada model *random forest* dan *XGBoost* memperoleh hasil nilai akurasi yang tinggi pada metode *Standard Scaler* dengan nilai masing-masing 0,86, sedangkan pada model *random forest* dan *XGBoost* dengan *Min Max Scaler* memperoleh hasil dengan nilai

masing-masing 0,85. Selain itu juga penelitian yang dilakukan oleh (Hartina Hiromi Satyanegara, 2022) dengan judul *Implementation of CNN-MLP and CNN-LSTM for MitM Attack Detection System* dengan hasil pada model CNN-MLP (*Convolutional Neural Network-Multilayer Perceptron*) dan CNN-LSTM (*Convolutional Neural Network-Long Short Term Memory*) dengan menggunakan metode *standard scaler* memperoleh nilai akurasi masing-masing sebesar 99,74% dan 99,44%. Sedangkan hasil yang didapatkan dengan metode *min-max scaler* memperoleh nilai akurasi masing-masing sebesar 99,67% dan 99,40%. Normalisasi yang dilakukan dalam penelitian ini dibagi menjadi 2 variabel yaitu variabel X dan variabel Y. Pemisahan variabel ini bertujuan untuk memudahkan langkah selanjutnya pada saat melakukan *training data* dan *testing data*. Hasil *feature scaling* dengan *standard scaler* dapat dilihat pada lampiran 25 dan 26 pada halaman C-9.

5.1.5 Analisis Hasil Matriks Korelasi

Matriks Korelasi atau *Correlation matrix* dalam penelitian ini dilakukan untuk mengetahui hubungan antar setiap variabel dalam dataset. *Heatmap* dalam *correlation matrix* digunakan untuk memudahkan dalam menemukan pola hubungan antar variabel, apabila terdapat warna terang dalam *heatmap* memiliki hubungan antar variabel yang tinggi dan sebaliknya. Selain itu juga dalam matriks terdapat angka, dimana angka yang mendekati angka 1 menunjukkan hubungan yang kuat antar variabel, angka -1 menandakan hubungan yang lemah antar variabel, dan angka 0 tidak menunjukkan adanya hubungan antar variabel. Berikut merupakan hasil visualisasi dari matriks korelasi:



Gambar 5. 76 Visualisasi Matriks Korelasi

Berdasarkan gambar 5.76, terdapat variabel yang memiliki nilai korelasi antar variabel mendekati 1, 0, dan -1. Nilai korelasi antar variabel dengan nilai mendekati 1 menandakan bahwa terdapat hubungan positif yang kuat. Hal ini dapat dilihat apabila terdapat nilai variabel yang meningkat, maka terdapat nilai variabel lain yang meningkat. Kemudian nilai korelasi antar variabel dengan nilai mendekati -1 menandakan adanya hubungan negatif atau lemah yang menandakan tidak terdapat korelasi terhadap antar variabel. Hal ini dapat dilihat apabila terdapat nilai variabel yang meningkat tetapi nilai variabel yang

lain menurun. Selanjutnya nilai korelasi antar variabel dengan nilai mendekati 0 menunjukkan bahwa variabel tersebut memiliki hubungan dimana tidak terdapat hubungan yang kuat maupun lemah. Nilai korelasi ini mengindikasikan tidak adanya hubungan saling keterkaitan pada variabel tersebut. Pada matriks korelasi terdapat *scale legend* atau skala legenda yang memberikan petunjuk mengenai hubungan antar variabel. Variabel dengan nilai 0,0 menandakan tidak terdapat korelasi pada variabel yang ditandai dengan warna kuning muda. Kemudian variabel dengan nilai 0,2 menandakan korelasi antar variabel memiliki korelasi yang lemah yang ditandai dengan warna hijau kekuningan. Selanjutnya variabel dengan nilai 0,4 menandakan korelasi antar variabel memiliki hubungan korelasi sedang yang ditandai dengan hijau nuda. Variabel dengan nilai 0,6 menunjukan hubungan antar variabel memiliki hubungan yang kuat antar variabel ditandai dengan warna hijau kebiruan. Variabel dengan nilai 0,8 menunjukan hubungan sangat kuat antar variabel yang ditandai dengan warna biru muda. Variabel dengan nilai 1,0 menujukan hubungan yang sempurna antar variabel dengan warna biru tua.

Dalam matriks korelasi terdapat variabel *feature* dan variabel target. Variabel *feature* digunakan sebagai masukan berupa nilai dalam melakukan prediksi yang disimbolkan dengan variabel X, sedangkan variabel target digunakan sebagai tujuan prediksi dalam kolom data yang disimbolkan dengan variabel Y. Berikut merupakan penjelasan mengenai hubungan antar variabel berdasarkan matriks korelasi pada gambar 5.10 :

5.1.5.1 Analisis Matriks Korelasi Variabel *Checkup*

Tabel 5. 3 Analisis Matriks Korelasi Variabel *Checkup*

Variabel *Feature*	**Variabel Target**
	General_Health
Checkup	0,027

Berdasarkan tabel 5.3 dapat diketahui bahwa hubungan variabel *checkup* dengan variabel *general_health* memiliki nilai sebesar 0,027. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *checkup* dan variabel

general health dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *checkup* dan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

5.1.5.2 Analisis Matriks Korelasi Variabel *Exercise*

Tabel 5. 4 Analisis Matriks Korelasi Variabel *Exercise*

Variabel *Feature*	**Variabel Target**	
	General_Health	*Checkup*
Exercise	0,037	-0,03

Berdasarkan tabel 5.4 dapat diketahui bahwa hubungan variabel *exercise* dengan variabel *general health* mendapatkan nilai 0,037. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *exercise* dan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *exercise* dan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan variabel *exercise* dengan variabel *checkup* mendapatkan nilai -0,03. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *exercise* dan variabel *checkup* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

5.1.5.3 Analisis Matriks Korelasi Variabel *Heart_Disease*

Tabel 5. 5 Analisis Matriks Korelasi Variabel *Heart Disease*

Variabel *Feature*	**Variabel Target**		
	General_Health	*Checkup*	*Exercise*
Heart_Disease	-0,021	0,084	-0,096

Berdasarkan tabel 5.5 dapat diketahui bahwa hubungan variabel *heart disease* dengan variabel *general health* mendapatkan nilai -0,021. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *heart disease* dan variabel *general health* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan variabel *heart disease* dengan variabel *checkup* mendapatkan nilai 0,084. Nilai tersebut menggambarkan hubungan dengan nilai positif.

Akan tetapi nilai hubungan antara variabel *heart disease* dengan variabel *checkup* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *heart disease* dengan variabel *checkup* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hbubungan antara variabel *heart disease* dengan variabel *exercise* mendapatkan nilai -0,096. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *heart disease* dengan variabel *exercise*, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

5.1.5.4 Analisis Matriks Korelasi Variabel *Skin Cancer*

Tabel 5. 6 Analisis Matriks Korelasi Variabel *Skin Cancer*

Variabel *Feature*	**Variabel Target**			
	General_Health	*Checkup*	*Exercise*	*Heart_Disease*
Skin_Cancer	0,021	0,079	-0,004	0,091

Berdasarkan tabel 5.6 dapat diketahui bahwa hubungan variabel *skin cancer* dengan variabel *general health* mendapatkan nilai 0,021. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *skin cancer* dengan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *skin cancer* dengan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan variabel *skin cancer* dengan variabel *checkup* mendapatkan nilai 0,079. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *skin cancer* dengan variabel *checkup* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *skin cancer* dengan variabel *checkup* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan antara variabel *skin cancer* dengan variabel *exercise* mendapatkan nilai -0,004. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *skin cancer* dengan variabel *exercise* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *skin cancer* dengan variabel *heart disease* mendapatkan nilai 0,091. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai

hubungan antara variabel *skin cancer* dengan variabel *heart disease* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *skin cancer* dengan variabel *heart disease* tidak memiliki pengaruh dan cenderung lemah.

5.1.5.5 Analisis Matriks Korelasi Variabel *Other Cancer*

Tabel 5. 7 Analisis Matriks Korelasi Variabel *Other Cancer*

Variabel Target / **Variabel *Feature***	*General_Health*	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Other_Cancer	0,0045	0,087	-0,054	0,092	0,15

Berdasarkan tabel 5.7 dapat diketahui bahwa hubungan variabel *other cancer* dengan variabel *general health* mendapatkan nilai 0,0045. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *other cancer* dan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *other cancer* dan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan variabel *other cancer* dengan variabel *checkup* mendapatkan nilai 0,087. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *other cancer* dan variabel *checkup* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *other cancer* dan variabel *checkup* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan antara variabel *other cancer* dengan variabel *exercise* mendapatkan nilai -0,054. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *other cancer* dengan variabel *exercise* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Hubungan variabel *other cancer* dengan variabel *heart disease* mendapatkan nilai 0,092. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *other cancer* dan variabel *heart disease* dengan nilai yang

mendekati 0. Sehingga hubungan antara variabel *other cancer* dan variabel *heart disease* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *other cancer* dengan variabel *skin cancer* mendapatkan nilai 0,15. Nilai tersebut menggambarkan hubungan dengan nilai positif. Hubungan variabel *other cancer* dengan variabel *skin cancer* memiliki nilai sekitar 0. Sehingga hubungan antara variabel *other cancer* dengan variabel *skin cancer* menandakan hubungan tersebut lemah dan cenderung tidak terdapat hubungan.

5.1.5.6 Analisis Matriks Korelasi Variabel *Depression*

Tabel 5. 8 Analisis Matriks Korelasi Variabel *Depression*

Variabel Target / **Variabel *Feature***	*General_Health*	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Depression	0,0023	0,036	-0,085	0,032	-0,013

Berdasarkan tabel 5.8 dapat diketahui bahwa hubungan variabel *depression* dengan variabel *general health* mendapatkan nilai 0,0023. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *depression* dan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *depression* dan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *depression* dengan variabel *checkup* mendapatkan nilai 0,036. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *depression* dan variabel *checkup* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *depression* dan variabel *checkup* tidak memiliki pengaruh dan cenderung lemah

Kemudian hubungan antara variabel *depression* dengan variabel *exercise* mendapatkan nilai -0,085. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *depression* dengan variabel *exercise* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Hubungan variabel *depression* dengan variabel *heart disease* mendapatkan nilai 0,032. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *depression* dan variabel *heart disease* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *depression* dan variabel *heart disease* tidak memiliki pengaruh dan cenderung lemah

Hubungan variabel *depression* dan variabel *skin cancer* mendapatkan nilai -0,013. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *depression* dan variabel *skin cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Tabel 5. 9 Analisis Matriks Korelasi Variabel *Depression*

Variabel *Feature*	**Variabel Target**
	Other_Cancer
Depression	0,016

Berdasarkan tabel 5.9 dapat diketahui bahwa hubungan variabel *depression* dengan variabel *other cancer* mendapatkan nilai 0,016. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *depression* dan variabel *other cancer* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *depression* dengan variabel *other cancer* tidak memiliki pengaruh dan cenderung lemah.

5.1.5.7 Analisis Matriks Korelasi Variabel *Diabetes*

Tabel 5. 10 Analisis Matriks Korelasi Variabel *Diabetes*

Variabel *Feature*	**Variabel Target**				
	General_Health	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Diabetes	-0,025	0,13	-0,14	0,17	-0,033

Berdasarkan tabel 5.10 dapat diketahui bahwa hubungan variabel *diabetes* dengan variabel *general health* mendapatkan nilai -0,025. Nilai tersebut menggambarkan

hubungan dengan nilai negatif. Hubungan antara variabel *diabetes* dan variabel *general health* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *diabetes* dengan variabel *checkup* mendapatkan nilai 0,013. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *diabetes* dan variabel *checkup* dengan nilai sekitar 0. Sehingga hubungan antara variabel *diabetes* dan variabel *checkup* cenderung lemah.

Hubungan antara variabel *diabetes* dengan variabel *exercise* mendapatkan nilai -0,14. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *diabetes* dengan variabel *exercise* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Hubungan variabel *diabetes* dengan variabel *heart disease* mendapatkan nilai 0,17. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *diabetes* dan variabel *heart disease* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *diabetes* dengan variabel *heart disease* memiliki hubungan cenderung lemah.

Hubungan variabel *diabetes* dengan variabel *skin cancer* mendapatkan nilai -0,033. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *diabetes* dengan variabel *skin cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Tabel 5. 11 Analisis Matriks Korelasi Variabel *Diabetes*

Variabel Target		
Variabel *Feature*		
	Other_Cancer	*Depression*
Diabetes	0,064	0,053

Berdasarkan tabel 5.11 dapat diketahui bahwa hubungan variabel *diabetes* dengan variabel *other cancer* mendapatkan nilai 0,064. Nilai tersebut menggambarkan hubungan

dengan nilai positif. Akan tetapi nilai hubungan antara variabel *diabetes* dan variabel *other cancer* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *diabetes* dan variabel *other cancer* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan variabel *diabetes* dengan variabel *checkup* mendapatkan nilai 0,053. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *diabetes* dan variabel *checkup* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *diabetes* dan variabel *checkup* tidak memiliki pengaruh dan cenderung lemah

5.1.5.8 Analisis Matriks Korelasi Variabel *Arthritis*

Tabel 5. 12 Analisis Matriks Korelasi Variabel *Arthritis*

Variabel Target / **Variabel *Feature***	*General_Health*	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Arthritis	0,014	0,15	-0,12	0,15	0,14

Berdasarkan tabel 5.12 dapat diketahui bahwa hubungan variabel *arthritis* dengan variabel *general health* mendapatkan nilai 0,014. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *arthritis* dan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *arthritis* dan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *arthritis* dan variabel *checkup* mendapatkan nilai 0,15. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *arthritis* dengan variabel *checkup* dengan nilai sekitar 0. Sehingga hubungan antara variabel *arthritis* dengan variabel *checkup* cenderung lemah.

Kemudian hubungan antara variabel *arthritis* dengan variabel *exercise* mendapatkan nilai -0,12. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *arthritis* dengan variabel *exercise* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *arthritis* dengan variabel *heart disease* mendapatkan nilai 0,15. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *arthritis* dengan variabel *heart disease* dengan nilai sekitar 0. Sehingga hubungan antara variabel *arthritis* dengan variabel *heart disease* cenderung lemah.

Hubungan variabel *arthritis* dan variabel *skin cancer* mendapatkan nilai 0,14. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *arthritis* dengan variabel *skin cancer* dengan nilai sekitar 0. Sehingga hubungan antara variabel *arthritis* dengan variabel *skin cancer* cenderung lemah.

Tabel 5. 13 Analisis Matriks Korelasi Variabel *Arthritis*

Variabel Target / **Variabel *Feature***	Other_Cancer	*Depression*	*Diabetes*
Arthritis	0,13	0,12	0,13

Berdasarkan tabel 5.13 dapat diketahui bahwa hubungan variabel *arthritis* dengan variabel *other cancer* mendapatkan nilai 0,13. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *arthritis* dan variabel *other cancer* dengan nilai sekitar 0. Sehingga hubungan antara variabel *arthritis* dan variabel *other cancer* cenderung lemah.

Kemudian hubungan variabel *arthritis* dengan variabel *depression* mendapatkan nilai 0,12. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *arthritis* dengan variabel *depression* dengan nilai sekitar 0. Sehingga hubungan antara variabel *arthritis* dengan variabel *depression* cenderung lemah.

Selanjutnya hubungan variabel *arthritis* dengan variabel *diabetes* mendapatkan nilai 0,13. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *arthritis* dengan variabel *diabetes* dengan nilai sekitar 0. Sehingga hubungan antara variabel *arthritis* dengan variabel *diabetes* cenderung lemah.

5.1.5.9 Analisis Matriks Korelasi Variabel *Sex*

Tabel 5. 14 Analisis Matriks Korelasi Variabel *Sex*

Variabel Target / **Variabel *Feature***	*General_Health*	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Sex	-0,014	-0,1	0,059	0,073	0,0097

Berdasarkan tabel 5.14 dapat diketahui bahwa hubungan variabel *sex* dan variabel *general health* mendapatkan nilai -0,014. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *sex* dengan variabel *general health* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *sex* dan variabel *checkup* mendapatkan nilai -0,1. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *sex* dan variabel *checkup* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan antara variabel *sex* dan variabel *exercise* mendapatkan nilai 0,059. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *sex* dan variabel *exercise* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *sex* dan variabel *exercise* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *sex* dengan variabel *heart disease* mendapatkan nilai 0,073. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *sex* dengan variabel *heart disease* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *sex* dengan variabel *heart disease* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *sex* dengan variabel *skin cancer* mendapatkan nilai 0,0097. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *sex* dengan variabel *skin cancer* dengan nilai yang mendekati 0. Sehingga

hubungan antara variabel *sex* dengan variabel *skin cancer* tidak memiliki pengaruh dan cenderung lemah.

Tabel 5. 15 Analisis Matriks Korelasi Variabel *Sex*

Variabel Target / **Variabel *Feature***				
	Other_Cancer	*Depression*	*Diabetes*	*Arthritis*
Sex	-0,042	-0,14	-0,013	-0,1

Berdasarkan tabel 5.15 dapat diketahui bahwa hubungan variabel *sex* dan variabel *other cancer* mendapatkan nilai -0,042. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *sex* dan variabel *other cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan variabel *sex* dan variabel *depression* mendapatkan nilai -0,14. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *sex* dan variabel *depression* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *sex* dengan variabel *diabetes* mendapatkan nilai -0,013. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *sex* dengan variabel *diabetes* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *sex* dan variabel *arthritis* mendapatkan nilai -0,1. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *sex* dan variabel *arthritis* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

5.1.5.10 Analisis Matriks Korelasi Variabel *Age Category*

Tabel 5. 16 Analisis Matriks Korelasi Variabel *Age Category*

Variabel Target / **Variabel *Feature***	*General_Health*	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Age_Category	0,032	0,22	-0,12	0,23	0,27

Berdasarkan tabel 5.16 dapat diketahui bahwa hubungan variabel *age category* dengan variabel *general health* mendapatkan nilai 0,032. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *age category* dengan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *age category* dengan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *age category* dengan variabel *checkup* mendapatkan nilai 0,22. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *age category* dengan variabel *checkup* dengan nilai yang sekitar 0. Sehingga hubungan antara variabel *age category* dengan variabel *checkup* memiliki hubungan yang lemah.

Kemudian hubungan antara variabel *age category* dengan variabel *exercise* mendapatkan nilai -0,12. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *age category* dengan variabel *exercise* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *age category* dengan variabel *heart disease* mendapatkan nilai 0,23. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *age category* dengan variabel *heart disease* dengan nilai yang sekitar 0. Sehingga hubungan antara variabel *age category* dengan variabel *heart disease* memiliki hubungan yang lemah.

Hubungan variabel *age category* dengan variabel *skin cancer* mendapatkan nilai 0,27. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai

hubungan antara variabel *age category* dengan variabel *skin cancer* dengan nilai yang sekitar 0. Sehingga hubungan antara variabel *age category* dengan variabel *skin cancer* memiliki hubungan yang lemah.

Tabel 5. 17 Analisis Matriks Korelasi Variabel *Age Category*

Variabel Target / **Variabel *Feature***	*Other_Cancer*	*Depression*	*Diabetes*	*Arthritis*	*Sex*
Age_Category	0,23	-0,1	0,19	0,37	-0,06

Berdasarkan tabel 5.17 dapat diketahui bahwa hubungan variabel *age category* dengan variabel *other cancer* mendapatkan nilai 0,23. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *age category* dengan variabel *other cancer* dengan nilai yang sekitar 0. Sehingga hubungan antara variabel *age category* dengan variabel *other cancer* memiliki hubungan yang lemah.

Kemudian hubungan variabel *age category* dengan variabel *depression* mendapatkan nilai -0,1. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *age category* dengan variabel *depression* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *age category* dengan variabel *diabetes* mendapatkan nilai 0,19. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *age category* dengan variabel *diabetes* dengan nilai yang sekitar 0. Sehingga hubungan antara variabel *age category* dengan variabel *diabetes* memiliki hubungan cenderung lemah.

Hubungan variabel *age category* dengan variabel *arthritis* mendapatkan nilai 0,37. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *age category* dengan variabel *arthritis* dengan nilai yang sekitar 0. Sehingga hubungan antara variabel *age category* dengan variabel *arthritis* memiliki hubungan yang lemah cenderung sedang.

Hubungan variabel *age category* dengan variabel *sex* mendapatkan nilai -0,06. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *age category* dengan variabel *sex* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

5.1.5.11 Analisis Matriks Korelasi Variabel *Height (cm)*

Tabel 5. 18 Analisis Matriks Korelasi Variabel *Height (cm)*

Variabel Target / **Variabel *Feature***					
	General_Health	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Height_(cm)	0,0016	-0,089	0,092	0,16	0,0068

Berdasarkan tabel 5.18 dapat diketahui bahwa hubungan variabel *height (cm)* dan variabel *general health* mendapatkan nilai 0,0016. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *height (cm)* dan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *height (cm)* dan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *height (cm)* dengan variabel *checkup* mendapatkan nilai -0,089. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *height (cm)* dengan variabel *checkup* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan antara variabel *height (cm)* dan variabel *exercise* mendapatkan nilai 0,092. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *height (cm)* dan variabel *exercise* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *height (cm)* dan variabel *exercise* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *height (cm)* dengan variabel *heart disease* mendapatkan nilai 0,16. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *height (cm)* dengan variabel *heart disease* dengan nilai yang

mendekati 0. Sehingga hubungan antara variabel *height (cm)* dengan variabel *heart disease* cenderung lemah.

Hubungan variabel *height (cm)* dengan variabel *skin cancer* mendapatkan nilai 0,0068. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *height (cm)* dengan variabel *skin cancer* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *height (cm)* dengan variabel *skin cancer* tidak memiliki pengaruh dan cenderung lemah.

Tabel 5. 19 Analisis Matriks Korelasi Variabel *Height (cm)*

Variabel Target / **Variabel *Feature***	Other_Cancer	*Depression*	*Diabetes*	*Arthritis*	*Sex*
Height_(cm)	-0,043	-0,091	-0,043	-0,091	0,7

Berdasarkan tabel 5.19 dapat diketahui bahwa hubungan variabel *height (cm)* dan variabel *other cancer* mendapatkan nilai -0,44. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *height (cm)* dan variabel *other cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *height (cm)* dan variabel *depression* mendapatkan nilai -0,092. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *height (cm)* dan variabel *depression* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan antara variabel *height (cm)* dan variabel *diabetes* mendapatkan nilai -0,043. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *height (cm)* dan variabel *diabetes* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *height (cm)* dan variabel *arthritis* mendapatkan nilai -0,091. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *height (cm)* dan variabel *arthritis* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *height (cm)* dan variabel *sex* mendapatkan nilai 0,7. Nilai tersebut menggambarkan hubungan dengan nilai positif. Nilai hubungan antara variabel *height (cm)* dan variabel *sex* dengan nilai yang mendekati 1. Sehingga hubungan antara variabel *height (cm)* dan variabel *sex* memiliki hubungan yang kuat.

Tabel 5. 20 Analisis Matriks Korelasi Variabel *Height (cm)*

Variabel Target	
Variabel *Feature*	
	Age_Category
Height_(cm)	-0,12

Berdasarkan tabel 5.20 dapat diketahui bahwa hubungan variabel *height (cm)* dan variabel *age category* mendapatkan nilai -0,12. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *height (cm)* dan variabel *age category* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

5.1.5.12 Analisis Matriks Korelasi Variabel *Weight (kg)*

Tabel 5. 21 Analisis Matriks Korelasi Variabel *Weight (kg)*

Variabel Target					
Variabel *Feature*					
	General_Health	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Weight_(kg)	0,0015	0,012	-0,09	0,046	-0,029

Berdasarkan tabel 5.21 dapat diketahui bahwa hubungan variabel *weight (kg)* dan variabel *general health* mendapatkan nilai 0,0015. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *weight (kg)* dan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *exercise* dan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan variabel *weight (kg)* dan variabel *checkup* mendapatkan nilai 0,012. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai

hubungan antara variabel *weight (kg)* dan variabel *checkup* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *weight (kg)* dan variabel *checkup* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan antara variabel *weight (kg)* dan variabel *exercise* mendapatkan nilai -0,09. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *weight (kg)* dan variabel *exercise* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *weight (kg)* dan variabel *heart disease* mendapatkan nilai 0,046. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *weight (kg)* dan variabel *heart disease* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *weight (kg)* dan variabel *heart disease* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *weight (kg)* dengan variabel *skin cancer* mendapatkan nilai -0,029. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *weight (kg)* dengan variabel *skin cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Tabel 5. 22 Analisis Matriks Korelasi Variabel *Weight (kg)*

Variabel Target **Variabel *Feature***					
	Other_Cancer	*Depression*	*Diabetes*	*Arthritis*	*Sex*
Weight_(kg)	-0,021	0,048	0,16	0,074	0,35

Berdasarkan tabel 5.22 dapat diketahui bahwa hubungan variabel *weight (kg)* dan variabel *other cancer* mendapatkan nilai -0,021. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *weight (kg)* dan variabel *other cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *weight (kg)* dan variabel *depression* mendapatkan nilai 0,048. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *weight (kg)* dan variabel *depression* dengan nilai yang

mendekati 0. Sehingga hubungan antara variabel *weight (kg)* dan variabel *depression* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan antara variabel *weight (kg)* dan variabel *diabetes* mendapatkan nilai 0,16. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *weight (kg)* dan variabel *diabetes* dengan nilai yang sekitar 0. Sehingga hubungan antara variabel *weight (kg)* dan variabel *diabetes* cenderung lemah.

Hubungan variabel *weight (kg)* dan variabel *arthritis* mendapatkan nilai 0,074. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *weight (kg)* dan variabel *arthritis* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *weight (kg)* dan variabel *arthritis* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *weight (kg)* dan variabel *sex* mendapatkan nilai 0,35 dengan warna hijau tua yang menandakan hubungan lemah. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *weight (kg)* dan variabel *sex* dengan nilai yang sekitar 0. Sehingga hubungan antara variabel *weight (kg)* dan variabel *sex* memiliki hubungan lemah yang cenderung sedang.

Tabel 5. 23 Analisis Matriks Korelasi Variabel *Weight (kg)*

Variabel Target		
Variabel *Feature*		
	Age_Category	*Height_(cm)*
Weight_(kg)	-0,062	0,47

Berdasarkan tabel 5.23 dapat diketahui bahwa hubungan variabel *weight (kg)* dan variabel *age category* mendapatkan nilai -0,062. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *weight (kg)* dan variabel *age category* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Kemudian variabel *weight (kg)* dan variabel *Height (cm)*mendapatkan nilai 0,47. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan

antara variabel *weight (kg)* dan variabel *Height (cm)* dengan nilai yang sekitar 0. Sehingga hubungan antara variabel *weight (kg)* dan variabel *Height (cm)* memiliki hubungan yang sedang.

5.1.5.13 Analisis Matriks Korelasi Variabel BMI

Tabel 5. 24 Analisis Matriks Korelasi Variabel *BMI*

Variabel Target / **Variabel *Feature***					
	General_Health	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
BMI	0,014	0,061	-0,16	0,043	-0,038

Berdasarkan tabel 5.24 dapat diketahui bahwa hubungan variabel *BMI* dan variabel *general health* mendapatkan nilai 0,0014. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *BMI* dan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *BMI* dan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan variabel *BMI* dan variabel *checkup* mendapatkan nilai 0,061. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *BMI* dan variabel *checkup* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *BMI* dan variabel *checkup* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan antara variabel *BMI* dan variabel *exercise* mendapatkan nilai -0,16. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *BMI* dan variabel *exercise* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *BMI* dengan variabel *heart disease* mendapatkan nilai 0,043. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *BMI* dengan variabel *heart disease* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *BMI* dengan variabel *heart disease* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *BMI* dengan variabel *skin cancer* mendapatkan nilai -0,038. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *BMI* dengan variabel *skin cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Tabel 5. 25 Analisis Matriks Korelasi Variabel *BMI*

Variabel Target / **Variabel *Feature***	*Other_Cancer*	*Depression*	*Diabetes*	*Arthritis*	*Sex*
BMI	0,001	0,11	0,21	0,14	0,011

Berdasarkan tabel 5.25 dapat diketahui bahwa hubungan variabel *BMI* dengan variabel *other cancer* mendapatkan nilai 0,001. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *BMI* dengan variabel *other cancer* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *BMI* dengan variabel *other cancer* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *BMI* dengan variabel *depression* mendapatkan nilai 0,11. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *BMI* dengan variabel *depression* dengan nilai berada disekitar 0. Sehingga hubungan antara variabel *BMI* dengan variabel *depression* cenderung lemah.

Kemudian hubungan antara variabel *BMI* dengan variabel *diabetes* mendapatkan nilai 0,21. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *BMI* dengan variabel *diabetes* dengan nilai berada disekitar 0. Sehingga hubungan antara variabel *BMI* dengan variabel *diabetes* cenderung lemah.

Hubungan variabel *BMI* dengan variabel *arthritis* mendapatkan nilai 0,14. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *BMI* dengan variabel *arthritis* dengan nilai berada disekitar 0. Sehingga hubungan antara variabel *BMI* dengan variabel *arthritis* cenderung lemah.

Hubungan variabel *BMI* dengan variabel *sex* mendapatkan nilai 0,011. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan

antara variabel *BMI* dengan variabel *sex* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *BMI* dengan variabel *sex* tidak memiliki pengaruh dan cenderung lemah.

Tabel 5. 26 Analisis Matriks Korelasi Variabel *BMI*

Variabel Target / **Variabel *Feature***	*Age_Category*	*Height_(cm)*	*Weight_(kg)*
BMI	-0,0074	-0,027	0,86

Berdasarkan tabel 5.26 dapat diketahui bahwa hubungan variabel *BMI* dan variabel *age category* mendapatkan nilai -0,0074. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *BMI* dan variabel *age category* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *BMI* dan variabel *height (cm)* mendapatkan nilai -0,027. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *BMI* dan variabel *height (cm)* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan antara variabel *BMI* dan variabel *weight (kg)* mendapatkan nilai 0,86. Nilai tersebut menggambarkan hubungan dengan nilai positif. Nilai hubungan antara variabel *BMI* dan variabel *weight (kg)* dengan nilai mendekati 1. Sehingga hubungan antara variabel *BMI* dan variabel *weight (kg)* memiliki hubungan yang sangat kuat.

5.1.5.14 Analisis Matriks Korelasi *Smoking History*

Tabel 5. 27 Analisis Matriks Korelasi Variabel *Smoking History*

Variabel Target / **Variabel *Feature***	*General_Health*	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Smoking_History	0,0043	-0,01	-0,093	0,11	0,033

Berdasarkan tabel 5.27 dapat diketahui bahwa hubungan variabel *smoking history* dengan variabel *general health* mendapatkan nilai 0,0043. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *smoking history* dengan variabel *checkup* mendapatkan nilai -0,01. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *smoking history* dengan variabel *checkup* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian Hubungan antara variabel *smoking history* dan variabel *exercise* mendapatkan nilai -0,093. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *smoking history* dan variabel *exercise* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *smoking history* dan variabel *heart disease* mendapatkan nilai 0,11. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dan variabel *heart disease* dengan nilai yang disekitar 0. Sehingga hubungan antara variabel *smoking history* dan variabel *heart disease* cenderung lemah.

Hubungan variabel *smoking history* dengan variabel *skin cancer* mendapatkan nilai 0,0033. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *skin cancer* dengan nilai

yang mendekati 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *skin cancer* tidak memiliki pengaruh dan cenderung lemah.

Tabel 5. 28 Analisis Matriks Korelasi Variabel *Smoking History*

Variabel Target / **Variabel *Feature***	*Other_Cancer*	*Depression*	*Diabetes*	*Arthritis*	*Sex*
Smoking_History	0,053	0,1	0,054	0,12	0,073

Berdasarkan tabel 5.28 dapat diketahui bahwa hubungan variabel *smoking history* dengan variabel *other cancer* mendapatkan nilai 0,053. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *other cancer* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *other cancer* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *smoking history* dengan variabel *depression* mendapatkan nilai 0,1. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *depression* dengan nilai berada disekitar 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *depression* cenderung lemah.

Kemudian hubungan antara variabel *smoking history* dengan variabel *diabetes* mendapatkan nilai 0,054. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *diabetes* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *diabetes* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *smoking history* dengan variabel *arthritis* mendapatkan nilai 0,12. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *arthritis* dengan nilai berada disekitar 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *arthritis* cenderung lemah.

Hubungan variabel *smoking history* dengan variabel *sex* mendapatkan nilai 0,073. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *sex* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *sex* tidak memiliki pengaruh dan cenderung lemah.

Tabel 5. 29 Analisis Matriks Korelasi Variabel *Smoking History*

Variabel Target **Variabel *Feature***				
	Age_Category	*Height_(cm)*	*Weight_(kg)*	*BMI*
Smoking_History	0,13	0,052	0,047	0,025

Berdasarkan tabel 5.29 dapat diketahui bahwa hubungan variabel *smoking history* dengan variabel *age category* mendapatkan nilai 0,13. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *age category* dengan nilai berada disekitar 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *age category* cenderung lemah.

Selanjutnya hubungan variabel *smoking history* dengan variabel *height (cm)* mendapatkan nilai 0,052. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *height (cm)* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *height (cm)* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan antara variabel *smoking history* dengan variabel *Weight (kg)* mendapatkan nilai 0,047. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *Weight (kg)* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *Weight (kg)* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *smoking history* dengan variabel *BMI* mendapatkan nilai 0,025. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *smoking history* dengan variabel *BMI* dengan nilai yang

mendekati 0. Sehingga hubungan antara variabel *smoking history* dengan variabel *BMI* tidak memiliki pengaruh dan cenderung lemah.

5.1.5.15 Analisis Matriks Korelasi Variabel *Alcohol Consumption*

Tabel 5. 30 Analisis Matriks Korelasi Variabel *Alcohol Consumption*

Variabel Target / **Variabel *Feature***	*General_Health*	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Alcohol_Consumption	0,0024	-0,047	0,095	-0,037	0,043

Berdasarkan tabel 5.30 dapat diketahui bahwa hubungan variabel *alcohol consumption* dengan variabel *general health* mendapatkan nilai 0,0024. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *alcohol consumption* dengan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *alcohol consumption* dengan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *alcohol consumption* dengan variabel *checkup* mendapatkan nilai -0,047. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *alcohol consumption* dengan variabel *checkup* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan antara variabel *alcohol consumption* dengan variabel *exercise* mendapatkan nilai 0,095. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *alcohol consumption* dengan variabel *exercise* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *alcohol consumption* dengan variabel *exercise* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *alcohol consumption* dan variabel *heart disease* mendapatkan nilai -0,037. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *alcohol consumption* dan variabel *heart disease* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *alcohol consumption* dengan variabel *skin cancer* mendapatkan nilai 0,043. Nilai tersebut menggambarkan hubungan dengan nilai positif.

Akan tetapi nilai hubungan antara variabel *alcohol consumption* dengan variabel *skin cancer* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *alcohol consumption* dengan variabel *skin cancer* tidak memiliki pengaruh dan cenderung lemah

Tabel 5. 31 Analisis Matriks Korelasi Variabel *Alcohol Consumption*

Variabel Target / **Variabel *Feature***	*Other_Cancer*	*Depression*	*Diabetes*	*Arthritis*	*Sex*
Alcohol_Consumption	-0,0087	-0,028	-0,11	-0,025	0,13

Berdasarkan tabel 5.31 dapat diketahui bahwa hubungan variabel *alcohol consumption* dan variabel *other cancer* mendapatkan nilai -0,0087. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *alcohol consumption* dan variabel *other cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *alcohol consumption* dengan variabel *depression* mendapatkan nilai -0,028. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *alcohol consumption* dengan variabel *depression* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan antara variabel *alcohol consumption* dengan variabel *diabetes* mendapatkan nilai -0,11. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *alcohol consumption* dengan variabel *diabetes* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *alcohol consumption* dan variabel *arthritis* mendapatkan nilai -0,025. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *alcohol consumption* dan variabel *arthritis* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *alcohol consumption* dan variabel *sex* mendapatkan nilai 0,013. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *alcohol consumption* dan variabel *sex* dengan nilai yang

mendekati 0. Sehingga hubungan antara variabel *alcohol consumption* dan variabel *sex* tidak memiliki pengaruh dan cenderung lemah

Tabel 5. 32 Analisis Matriks Korelasi Variabel *Alcohol Consumption*

Variabel Target / **Variabel *Feature***	*Age_Category*	*Height_(cm)*	*Weight_(kg)*	*BMI*	*Smoking_History*
Alcohol_Consumption	0,013	0,13	-0,032	-0,11	0,1

Berdasarkan tabel 5.32 dapat diketahui bahwa hubungan variabel *alcohol consumption* dan variabel *age category* mendapatkan nilai 0,013. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *alcohol consumption* dan variabel *age category* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *alcohol consumption* dan variabel *age category* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *alcohol consumption* dengan variabel *height (cm)* mendapatkan nilai 0,13. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *alcohol consumption* dengan variabel *height (cm)* dengan nilai yang disekitar 0. Sehingga hubungan antara variabel *alcohol consumption* dengan variabel *height (cm)* cenderung lemah.

Kemudian hubungan antara variabel *alcohol consumption* dengan variabel *Weight (kg)* mendapatkan nilai -0,032. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *alcohol consumption* dengan variabel *Weight (kg)* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *alcohol consumption* dengan variabel *BMI* mendapatkan nilai -0,11. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *alcohol consumption* dengan variabel *BMI* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *alcohol consumption* dengan variabel *smoking history* mendapatkan nilai 0,1. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *alcohol consumption* dengan variabel *smoking*

history dengan nilai yang disekitar 0. Sehingga hubungan antara variabel *alcohol consumption* dengan variabel *smoking history* cenderung lemah.

5.1.5.16 Analisis Matriks Korelasi Variabel *Fruit Consumption*

Tabel 5. 33 Analisis Matriks Korelasi Variabel *Fruit Consumption*

Variabel Target / **Variabel *Feature***	*General_Health*	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Fruit_Consumption	0,0011	0,039	0,14	-0,024	0,013

Berdasarkan tabel 5.33 dapat diketahui bahwa hubungan variabel *fruit consumption* dengan variabel *general health* mendapatkan nilai 0,0011. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fruit consumption* dengan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *fruit consumption* dengan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *fruit consumption* dengan variabel *checkup* mendapatkan nilai 0,039. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fruit consumption* dengan variabel *checkup* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *fruit consumption* dengan variabel *checkup* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan antara variabel *fruit consumption* dengan variabel *exercise* mendapatkan nilai 0,14. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fruit consumption* dengan variabel *exercise* dengan nilai yang disekitar 0. Sehingga hubungan antara variabel *fruit consumption* dengan variabel *exercise* cenderung lemah.

Hubungan variabel *fruit consumption* dengan variabel *heart disease* mendapatkan nilai -0,024. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fruit consumption* dengan variabel *heart disease* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Hubungan variabel *fruit consumption* dengan variabel *skin cancer* mendapatkan nilai 0,013. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fruit consumption* dengan variabel *skin cancer* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *fruit consumption* dengan variabel *skin cancer* tidak memiliki pengaruh dan cenderung lemah.

Tabel 5. 34 Analisis Matriks Korelasi Variabel *Fruit Consumption*

Variabel Target / **Variabel *Feature***	*Other_Cancer*	*Depression*	*Diabetes*	*Arthritis*	*Sex*
Fruit_Consumption	0,008	-0,04	-0,02	-0,0019	-0,002

Berdasarkan tabel 5.34 Dapat diketahui bahwa hubungan variabel *fruit consumption* dengan variabel *other cancer* mendapatkan nilai 0,008. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fruit consumption* dengan variabel *other cancer* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *fruit consumption* dengan variabel *other cancer* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *fruit consumption* dengan variabel *depression* mendapatkan nilai -0,04. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fruit consumption* dengan variabel *depression* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan antara variabel *fruit consumption* dengan variabel *diabetes* mendapatkan nilai -0,02. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fruit consumption* dengan variabel *diabetes* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *fruit consumption* dengan variabel *arthritis* mendapatkan nilai -0,0019. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fruit consumption* dengan variabel *arthritis* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *fruit consumption* dengan variabel *sex* mendapatkan nilai -0,002. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fruit consumption* dengan variabel *sex* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Tabel 5. 35 Analisis Matriks Korelasi Variabel *Fruit Consumption*

Variabel Target / **Variabel *Feature***	*Age_Category*	*Height_(cm)*	*Weight_(kg)*	*BMI*	*Smoking_History*
Fruit_Consumption	0,044	-0,046	-0,091	-0,077	-0,094

Berdasarkan tabel 5.35 dapat diketahui bahwa hubungan variabel *fruit consumption* dengan variabel *age category* mendapatkan nilai 0,044. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fruit consumption* dengan variabel *age category* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *fruit consumption* dengan variabel *age category* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *fruit consumption* dengan variabel *height (cm)* mendapatkan nilai -0,046. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fruit consumption* dengan variabel *height (cm)* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan antara variabel *fruit consumption* dengan variabel *Weight (kg)* mendapatkan nilai -0,097. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fruit consumption* dengan variabel *Weight (kg)* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *fruit consumption* dan variabel *BMI* mendapatkan nilai -0,077. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fruit consumption* dan variabel *BMI* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *fruit consumption* dengan variabel *smoking history* mendapatkan nilai -0,094. Nilai tersebut menggambarkan hubungan dengan nilai negatif.

Hubungan antara variabel *fruit consumption* dengan variabel *smoking history* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Tabel 5. 36 Analisis Matriks Korelasi Variabel *Fruit Consumption*

Variabel *Feature*	**Variabel Target**
	Alcohol_Consumption
Fruit_Consumption	-0,013

Berdasarkan tabel 5.36 dapat diketahui hubungan variabel *fruit consumption* dengan variabel *alcohol consumption* mendapatkan nilai -0,013. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fruit consumption* dengan variabel *alcohol consumption* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

5.1.5.17 Analisis Matriks Korelasi Variabel *Green Vegetable Consumption*

Tabel 5. 37 Analisis Matriks Korelasi Variabel *Green Vegetable Consumption*

Variabel *Feature*	**Variabel Target**				
	General_Health	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
Green_Vegetable_Consumption	-0,022	0,035	0,12	-0,024	0,013

Berdasarkan tabel 5.37 dapat diketahui bahwa hubungan variabel *Green Vegetable Consumption* dengan variabel *general health* mendapatkan nilai -0,022. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dengan variabel *general health* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *Green Vegetable Consumption* dengan variabel *checkup* mendapatkan nilai 0,035. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *Green Vegetable Consumption* dengan variabel *checkup* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel

Green Vegetable Consumption dengan variabel *checkup* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan antara variabel *Green Vegetable Consumption* dengan variabel *exercise* mendapatkan nilai 0,12. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *Green Vegetable Consumption* dengan variabel *exercise* dengan nilai yang disekitar 0. Sehingga hubungan antara variabel *Green Vegetable Consumption* dengan variabel *exercise* cenderung lemah.

Hubungan variabel *Green Vegetable Consumption* dengan variabel *heart disease* mendapatkan nilai -0,024. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dengan variabel *heart disease* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Hubungan variabel *Green Vegetable Consumption* dengan variabel *skin cancer* mendapatkan nilai 0,013. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *Green Vegetable Consumption* dengan variabel *skin cancer* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *Green Vegetable Consumption* dengan variabel *skin cancer* tidak memiliki pengaruh dan cenderung lemah.

Tabel 5. 38 Analisis Matriks Korelasi Variabel *Green Vegetable Consumption*

Variabel Target **Variabel *Feature***					
	Other_Cancer	*Depression*	*Diabetes*	*Arthritis*	*Sex*
Green_Vegetable _Consumption	-0,0032	-0,051	-0,027	-0,019	-0,069

Berdasarkan tabel 5.38 dapat diketahui bahwa hubungan variabel *Green Vegetable Consumption* dan variabel *other cancer* mendapatkan nilai -0,0032. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dan variabel *other cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *Green Vegetable Consumption* dan variabel *depression* mendapatkan nilai -0,051. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dan variabel *depression* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan antara variabel *Green Vegetable Consumption* dengan variabel *diabetes* mendapatkan nilai -0,027. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dengan variabel *diabetes* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *Green Vegetable Consumption* dengan variabel *arthritis* mendapatkan nilai -0,019. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dengan variabel *arthritis* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *Green Vegetable Consumption* dengan variabel *sex* mendapatkan nilai -0,069. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dengan variabel *sex* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Tabel 5. 39 Analisis Matriks Korelasi Variabel *Green Vegetable Consumption*

Variabel Target / **Variabel *Feature***					
	Age_Category	*Height_(cm)*	*Weight_(kg)*	*BMI*	*Smoking_History*
Green_Vegetable_Consumption	0,036	-0,03	-0,076	-0,071	-0,034

Berdasarkan tabel 5.39 dapat diketahui bahwa hubungan variabel *Green Vegetable Consumption* dengan variabel *age category* mendapatkan nilai 0,036. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *Green Vegetable Consumption* dengan variabel *age category* dengan nilai yang

mendekati 0. Sehingga hubungan antara variabel *Green Vegetable Consumption* dengan variabel *age category* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *Green Vegetable Consumption* dan variabel *height (cm)* mendapatkan nilai -0,03. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dan variabel *height (cm)* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Kemudian hubungan antara variabel *Green Vegetable Consumption* dan variabel *Weight (kg)* mendapatkan nilai -0,076. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dan variabel *Weight (kg)* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *Green Vegetable Consumption* dengan variabel *BMI* mendapatkan nilai -0,071. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dengan variabel *BMI* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *Green Vegetable Consumption* dengan variabel *smoking history* mendapatkan nilai -0,034. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *Green Vegetable Consumption* dengan variabel *smoking history* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Tabel 5. 40 Analisis Matriks Korelasi Variabel *Green Vegetable Consumption*

	Variabel Target	
Variabel *Feature*		
	Alcohol_Consumption	*Fruit_Consumption*
Green_Vegetable _Consumption	0,06	0,27

Berdasarkan tabel 5.40 dapat diketahui bahwa hubungan variabel *Green Vegetable Consumption* dengan variabel *alcohol consumption* mendapatkan nilai 0,06. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *Green Vegetable Consumption* dengan variabel *alcohol consumption* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *Green Vegetable Consumption* dengan variabel *alcohol consumption* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *Green Vegetable Consumption* dan variabel *Fruit Consumption* mendapatkan nilai 0,27. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *Green Vegetable Consumption* dan variabel *Fruit Consumption* dengan nilai disekitar 0. Sehingga hubungan antara variabel *Green Vegetable Consumption* dan variabel *Fruit Consumption* memiliki hubungan yang lemah.

5.1.5.18 Analisis Matriks Korelasi Variabel *Fried Potato Consumption*

Tabel 5. 41 Analisis Matriks Korelasi Variabel *Fried Potato Consumption*

Variabel Target / **Variabel *Feature***					
	General_Health	*Checkup*	*Exercise*	*Heart_Disease*	*Skin_Cancer*
FriedPotato _Consumption	0,00068	-0,056	-0,037	-0,0092	-0,039

Berdasarkan tabel 5.41 dapat diketahui bahwa hubungan variabel *fried potato consumption* dengan variabel *general health* mendapatkan nilai 0,0068. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fried potato consumption* dengan variabel *general health* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *fried potato consumption* dengan variabel *general health* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *fried potato consumption* dengan variabel *checkup* mendapatkan nilai -0,056. Nilai tersebut menggambarkan hubungan dengan nilai negatif.

Hubungan antara variabel *fried potato consumption* dengan variabel *checkup* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Kemudian hubungan antara variabel *fried potato consumption* dan variabel *exercise* mendapatkan nilai -0,037. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fried potato consumption* dan variabel *exercise* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *fried potato consumption* dan variabel *heart disease* mendapatkan nilai -0,092. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fried potato consumption* dan variabel *heart disease* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *fried potato consumption* dengan variabel *skin cancer* mendapatkan nilai -0,039. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fried potato consumption* dengan variabel *skin cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Tabel 5. 42 Analisis Matriks Korelasi Variabel *Fried Potato Consumption*

Variabel Target / **Variabel *Feature***	*Other_Cancer*	*Depression*	*Diabetes*	*Arthritis*	*Sex*
FriedPotato_Consumption_Consumption	-0,033	0,018	-0,0043	-0,051	0,013

Berdasarkan tabel 5.42 dapat diketahui bahwa hubungan variabel *fried potato consumption* dengan variabel *other cancer* mendapatkan nilai -0,033. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fried potato consumption* dengan variabel *other cancer* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *fried potato consumption* dan variabel *depression* mendapatkan nilai 0,018. Nilai tersebut menggambarkan hubungan dengan nilai positif.

Akan tetapi nilai hubungan antara variabel *fried potato consumption* dan variabel *depression* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *fried potato consumption* dan variabel *depression* tidak memiliki pengaruh dan cenderung lemah.

Kemudian hubungan antara variabel *fried potato consumption* dan variabel *diabetes* mendapatkan nilai -0,0043. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fried potato consumption* dan variabel *diabetes* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *fried potato consumption* dengan variabel *arthritis* mendapatkan nilai -0,051. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fried potato consumption* dengan variabel *arthritis* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Hubungan variabel *fried potato consumption* dan variabel *sex* mendapatkan nilai 0,013. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fried potato consumption* dan variabel *sex* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *fried potato consumption* dan variabel *sex* tidak memiliki pengaruh dan cenderung lemah.

Tabel 5. 43 Analisis Matriks Korelasi Variabel *Fried Potato Consumption*

Variabel Target / **Variabel *Feature***					
	Age_Category	*Height_(cm)*	*Weight_(kg)*	*BMI*	*Smoking_History*
FriedPotato_Consumption	-0,14	0,11	0,096	0,048	0,036

Berdasarkan tabel 5.43 dapat diketahui bahwa hubungan variabel *fried potato consumption* dan variabel *age category* mendapatkan nilai -0,14. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fried potato consumption* dan variabel *age category* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut.

Selanjutnya hubungan variabel *fried potato consumption* dengan variabel *height (cm)* mendapatkan nilai 0,11. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fried potato consumption* dengan variabel *height (cm)* dengan nilai yang disekitar 0. Sehingga hubungan antara variabel *fried potato consumption* dengan variabel *height (cm)* cenderung lemah.

Kemudian hubungan antara variabel *fried potato consumption* dan variabel *Weight (kg)* mendapatkan nilai 0,096. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fried potato consumption* dan variabel *Weight (kg)* dengan nilai yang disekitar 0. Sehingga hubungan antara variabel *fried potato consumption* dan variabel *Weight (kg)* cenderung lemah.

Hubungan variabel *fried potato consumption* dengan variabel *BMI* mendapatkan nilai 0,048. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fried potato consumption* dengan variabel *BMI* dengan nilai yang disekitar 0. Sehingga hubungan antara variabel *fried potato consumption* dengan variabel *BMI* tidak memiliki pengaruh dan cenderung lemah.

Hubungan variabel *fried potato consumption* dengan variabel *smoking history* mendapatkan nilai 0,036. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fried potato consumption* dengan variabel *smoking history* dengan nilai yang disekitar 0. Sehingga hubungan antara variabel *fried potato consumption* dengan variabel *smoking history* tidak memiliki pengaruh dan cenderung lemah.

Tabel 5. 44 Analisis Matriks Korelasi Variabel *Fried Potato Consumption*

	Variabel Target		
Variabel *Feature*			
	Alcohol_Consumption	*Fruit_Consumption*	*Green_Vegetable_Consumption*
FriedPotato_Consumption	0,021	-0,06	0,0032

Berdasarkan tabel 5.44 dapat diketahui bahwa hubungan variabel *fried potato consumption* dengan variabel *alcohol consumption* mendapatkan nilai 0,021. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fried potato consumption* dengan variabel *alcohol consumption* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *fried potato consumption* dengan variabel *alcohol consumption* tidak memiliki pengaruh dan cenderung lemah.

Selanjutnya hubungan variabel *fried potato consumption* dan variabel *fruit consumption* mendapatkan nilai -0,06. Nilai tersebut menggambarkan hubungan dengan nilai negatif. Hubungan antara variabel *fried potato consumption* dan variabel *fruit consumption* memiliki nilai negatif, sehingga tidak terdapat hubungan diantara dua variabel tersebut

Kemudian hubungan antara variabel *fried potato consumption* dengan variabel *green vegetable consumption* mendapatkan nilai 0,0032. Nilai tersebut menggambarkan hubungan dengan nilai positif. Akan tetapi nilai hubungan antara variabel *fried potato consumption* dengan variabel *green vegetable consumption* dengan nilai yang mendekati 0. Sehingga hubungan antara variabel *fried potato consumption* dengan variabel *green vegetable consumption* tidak memiliki pengaruh dan cenderung lemah.

5.1.6 Analisis Hasil Deskripsi Data

Deskripsi data atau *data describe* dilakukan untuk mengetahui gambaran statistik dalam dataset secara ringkas pada data numerik. Berikut merupakan hasil deskripsi data :

	count	mean	std	min	25%	50%	75%	max
General_Health	308774.0	2.273232	1.494016	0.00	1.00	2.00	4.00	4.00
Checkup	308774.0	3.514975	1.019649	0.00	4.00	4.00	4.00	4.00
Exercise	308774.0	0.775017	0.417572	0.00	1.00	1.00	1.00	1.00
Heart_Disease	308774.0	0.080871	0.272638	0.00	0.00	0.00	0.00	1.00
Skin_Cancer	308774.0	0.097133	0.296139	0.00	0.00	0.00	0.00	1.00
Other_Cancer	308774.0	0.096760	0.295631	0.00	0.00	0.00	0.00	1.00
Depression	308774.0	0.200467	0.400350	0.00	0.00	0.00	0.00	1.00
Diabetes	308774.0	0.308232	0.724454	0.00	0.00	0.00	0.00	3.00
Arthritis	308774.0	0.327304	0.469230	0.00	0.00	0.00	1.00	1.00
Sex	308774.0	0.481320	0.499652	0.00	0.00	0.00	1.00	1.00
Age_Category	308774.0	6.536104	3.523495	0.00	4.00	7.00	9.00	12.00
Height_(cm)	308774.0	170.615220	10.658452	91.00	163.00	170.00	178.00	241.00
Weight_(kg)	308774.0	83.590399	21.344664	24.95	68.04	81.65	95.25	293.02
BMI	308774.0	28.626813	6.522810	12.02	24.21	27.44	31.85	99.33
Smoking_History	308774.0	0.405662	0.491021	0.00	0.00	0.00	1.00	1.00
Alcohol_Consumption	308774.0	5.097557	8.200434	0.00	0.00	1.00	6.00	30.00
Fruit_Consumption	308774.0	29.834290	24.877812	0.00	12.00	30.00	30.00	120.00
Green_Vegetables_Consumption	308774.0	15.109517	14.926912	0.00	4.00	12.00	20.00	128.00
FriedPotato_Consumption	308774.0	6.297237	8.583837	0.00	2.00	4.00	8.00	128.00

Gambar 5. 77 Hasil Deskripsi Data

5.1.6.1 Analisis Deskripsi Data Variabel *General Health*

Tabel 5. 45 Analisis Deskripsi Data Variabel *General Health*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
General_Health	308.774	2.273232	1.494016	0.0	1.0	2.0	4.0	4.0

Berdasarkan tabel 5.45 Variabel *general_health* terdiri dari 308.774 yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 dan nilai maksimal atau terbesar dalam variabel yaitu 4. Nilai *mean* pada variabel ini yaitu 2.273232 dimana

menggambarkan kondisi kesehatan secara umum di variabel ini di ilustrasikan dalam keadaan rata-rata yang berada ditengah-tengah antara nilai minimal dan nilai maksimal. Nilai standar deviasi pada variabel ini yaitu 1.494016 yang menggambarkan pola persebaran data pada variabel ini. Nilai standar deviasi pada variabel ini dibawah nilai *mean*, sehingga persebaran dan keragaman data yang terdapat pada variabel ini kurang beragam dan tersebar. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) memiliki nilai 1 yang artinya *fair*, nilai kuartil 2 (50%) atau median memiliki nilai 2 yang artinya *good*, dan nilai kuartil 3 (75%) yaitu 4 yang artinya *very good*. Sehingga pada variabel ini dapat dikatakan bahwa kondisi kesehatan secara umum pada masyarakat sudah memiliki nilai yang baik.

5.1.6.2 Analisis Deskripsi Data Variabel *Checkup*

Tabel 5. 46 Analisis Deskripsi Data Variabel *Checkup*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
Checkup	308.774	3.514975	1.019649	0.0	4.0	4.0	4.0	4.0

Berdasarkan tabel 5.46 Variabel *checkup* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 dan nilai maksimal atau terbesar dalam variabel yaitu 4. Nilai *mean* pada variabel ini yaitu 3.514975 dimana menggambarkan waktu pemeriksaan kesehatan rutin yang dilakukan oleh masyarakat telah dilakukan secara rutin dengan nilai minimal atau terkecil pada variabel ini yaitu 0 dan nilai maksimal atau terbesar pada variabel ini yaitu 4. Nilai standar deviasi pada variabel ini yaitu 1.019649 yang memberikan gambaran pola persebaran data pada variabel ini. Nilai standar deviasi pada variabel ini berada dibawah nilai *mean*, sehingga persebaran serta keragaman data pada variabel ini kurang tersebar dan beragam. Akan tetapi korelasi nilai standar deviasi dan *mean* menunjukan nilai yang terdapat pada variabel ini memiliki nilai yang dekat dengan nilai rata-rata dengan jumlah persebaran yang terbatas. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%), nilai kuartil 2 (50%) atau median, dan nilai kuartil 3 (75%) memiliki nilai masing-masing 4 yang artinya *within the past year* atau pemeriksaan kesehatan yang dilakukan oleh masyarakat rutin dilakukan setiap tahun. Sehingga pada variabel ini dapat dikatakan bahwa masyarakat melakukan pemeriksaan kesehatan secara rutin yang secara umum dilakukan setiap tahun.

5.1.6.3 Analisis Deskripsi Data Variabel *Exercise*

Tabel 5. 47 Analisis Deskripsi Data Variabel *Exercise*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
Exercise	308.774	0.775017	0.417572	0.0	1.0	1.0	1.0	1.0

Berdasarkan tabel 5. 47 Variabel *checkup* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 dan nilai maksimal atau terbesar dalam variabel yaitu 1. Nilai *mean* pada variabel ini yaitu 0.775017 dimana menggambarkan masyarakat melakukan kegiatan aktivitas fisik dimana banyak masyarakat yang melakukan aktivitas yang dapat dilihat pada nilai *mean* yang mendekati 1. Nilai 1 ini artinya menunjukan masyarakat melakukan aktivitas fisik dibandingkan dengan nilai 0 yang artinya masyarakat tidak melakukan aktivitas fisik. Nilai standar deviasi pada variabel ini yaitu 0.417572 yang memberikan gambaran pola persebaran data pada variabel ini. Nilai standar deviasi pada variabel ini berada dibawah nilai *mean*, sehingga persebaran serta keragaman data pada variabel ini kurang tersebar dan beragam. Nilai *mean* dan nilai standar deviasi pada penelitian ini memiliki korelasi yang menunjukan variasi data pada variabel ini kurang bervariasi dan persebaran data yang kurang luas. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%), nilai kuartil 2 (50%) atau median, dan nilai kuartil 3 (75%) memiliki nilai masing-masing 1 yang artinya *yes* atau masyarakat melakukan kegiatan aktivitas fisik. Sehingga variabel ini dapat dikatakan bahwa masyarakat melakukan kegiatan aktivitas fisik.

5.1.6.4 Analisis Deskripsi Data Variabel *Heart Disease*

Tabel 5. 48 Analisis Deskripsi Data Variabel *Heart Disease*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
Heart_Disease	308.774	0.080871	0.272638	0.0	0.0	0.0	0.0	1.0

Berdasarkan tabel 5. 48 Variabel *heart disease* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya tidak terkena penyakit jantung dan nilai maksimal atau terbesar dalam variabel yaitu 1 yang artinya terkena penyakit jantung. Nilai *mean* pada variabel ini yaitu 0.080871 dimana menggambarkan masyarakat tidak menderita penyakit jantung. Nilai standar deviasi pada variabel ini yaitu 0.272638 yang menggambarkan pola persebaran data variabel ini. Nilai

standar deviasi pada variabel memiliki nilai yang lebih besar dibandingkan dengan nilai *mean* yang menunjukan pola persebaran data yang lebih beragam. Nilai *mean* dengan nilai 0,08 menunjukan bahwa nilai data yang terdapat pada variabel menandakan nilai mendekati dengan 0 yang artinya masyarakat tidak menderita penyakit jantung. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%), nilai kuartil 2 (50%) atau median, dan nilai kuartil 3 (75%) memiliki nilai masing-masing 0 yang artinya *no* atau masyarakat tidak menderita penyakitjantung. Sehingga variabel ini dapat dikatakan bahwa masyarakat secara umum masyarakat tidak menderita penyakit jantung.

5.1.6.5 Analisis Deskripsi Data Variabel *Skin Cancer*

Tabel 5. 49 Analisis Deskripsi Data Variabel *Skin Cancer*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
Skin_Cancer	308.774	0.097133	0.296139	0.0	0.0	0.0	0.0	1.0

Berdasarkan tabel 5.49 Variabel *skin cancer* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya tidak mengalami penyakit kanker kulit dan nilai maksimal atau terbesar dalam variabel yaitu 1 yang artinya mengalami penyakit kanker kulit. Nilai *mean* pada variabel ini yaitu 0.097133 dimana menggambarkan masyarakat tidak menderita penyakit kanker kulit. Nilai standar deviasi pada variabel ini yaitu 0.296139 yang menggambarkan pola persebaran data variabel ini. Nilai standar deviasi pada variabel memiliki nilai yang lebih besar dibandingkan dengan nilai *mean* yang menunjukan pola persebaran data yang lebih beragam. Nilai *mean* dengan nilai 0,097 menunjukan bahwa nilai data yang terdapat pada variabel menandakan nilai mendekati dengan 0 yang artinya masyarakat tidak menderita penyakit kanker kulit. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%), nilai kuartil 2 (50%) atau median, dan nilai kuartil 3 (75%) memiliki nilai masing-masing 0 yang artinya *no* atau masyarakat tidak menderita penyakit kanker kulit. Sehingga variabel ini dapat dikatakan bahwa masyarakat secara umum masyarakat tidak menderita penyakit kanker kulit.

5.1.6.6 Analisis Deskripsi Data Variabel *Other Cancer*

Tabel 5. 50 Analisis Deskripsi Data Variabel *Other Cancer*

Variabel	*Count*	*Mean*	*Std*	*Min*	*25%*	*50%*	*75%*	*Max*
Other_Cancer	308.774	0.096760	0.295631	0.0	0.0	0.0	0.0	1.0

Berdasarkan tabel 5.50 Variabel *other cancer* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya tidak mengalami penyakit kanker dengan jenis yang lain dan nilai maksimal atau terbesar dalam variabel yaitu 1 yang artinya mengalami penyakit kanker dengan jenis yang lain. Nilai *mean* pada variabel ini yaitu 0.096760 dimana menggambarkan masyarakat tidak menderita penyakit kanker dengan jenis yang lain. Nilai standar deviasi pada variabel ini yaitu 0.295631 yang menggambarkan pola persebaran data variabel ini. Nilai standar deviasi pada variabel memiliki nilai yang lebih besar dibandingkan dengan nilai *mean* yang menunjukan pola persebaran data yang lebih beragam. Nilai *mean* dengan nilai 0,096 menunjukan bahwa nilai data yang terdapat pada variabel menandakan nilai mendekati dengan 0 yang artinya masyarakat tidak menderita penyakit kanker dengan jenis yang lain. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%), nilai kuartil 2 (50%) atau median, dan nilai kuartil 3 (75%) memiliki nilai masing-masing 0 yang artinya *no* atau masyarakat tidak menderita penyakit kanker dengan jenis yang lain. Sehingga variabel ini dapat dikatakan bahwa masyarakat secara umum masyarakat tidak menderita penyakit kanker dengan jenis yang lain selain kanker kulit.

5.1.6.7 Analisis Deskripsi Data Variabel *Depression*

Tabel 5. 51 Analisis Deskripsi Data Variabel *Depression*

Variabel	*Count*	*Mean*	*Std*	*Min*	*25%*	*50%*	*75%*	*Max*
Depression	308.774	0.200467	0.400350	0.0	0.0	0.0	0.0	1.0

Berdasarkan tabel 5.51 Variabel *depression* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya tidak mengalami penyakit depresi dan nilai maksimal atau terbesar dalam variabel yaitu 1 yang artinya mengalami penyakit depresi. Nilai *mean* pada variabel ini yaitu 0.200467 dimana menggambarkan masyarakat tidak menderita penyakit depresi. Nilai standar deviasi pada variabel ini yaitu 0.400350 yang menggambarkan pola persebaran data variabel ini. Nilai

standar deviasi pada variabel memiliki nilai yang lebih besar dibandingkan dengan nilai *mean* yang menunjukan pola persebaran data yang lebih beragam. Nilai *mean* dengan nilai 0,200 menunjukan bahwa nilai data yang terdapat pada variabel menandakan nilai mendekati dengan 0 yang artinya masyarakat tidak menderita penyakit depresi. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%), nilai kuartil 2 (50%) atau median, dan nilai kuartil 3 (75%) memiliki nilai masing-masing 0 yang artinya *no* atau masyarakat tidak menderita penyakit depresi. Sehingga variabel ini dapat dikatakan bahwa masyarakat secara umum masyarakat tidak menderita penyakit depresi.

5.1.6.8 Analisis Deskripsi Data Variabel *Diabetes*

Tabel 5. 52 Analisis Deskripsi Data Variabel *Diabetes*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
Diabetes	308.774	0.308232	0.724454	0.0	0.0	0.0	0.0	3.0

Berdasarkan tabel 5.52 Variabel *diabetes* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya tidak mengalami penyakit depresi dan nilai maksimal atau terbesar dalam variabel yaitu 3 yang artinya *yes-but female told only during pregnancy* yang artinya ya, perempuan menderita penyakit diabetes tetapi diberitahu ketika kondisi selama proses kehamilan. Nilai *mean* pada variabel ini yaitu 0.308232 dimana menggambarkan masyarakat tidak menderita penyakit diabetes. Nilai standar deviasi pada variabel ini yaitu 0.724454 yang menggambarkan pola persebaran data variabel ini. Nilai standar deviasi pada variabel memiliki nilai yang lebih besar dibandingkan dengan nilai *mean* yang menunjukan pola persebaran data yang lebih beragam. Nilai *mean* dengan nilai 0,308 menunjukan bahwa nilai data yang terdapat pada variabel menandakan nilai mendekati dengan 0 yang artinya masyarakat tidak menderita penyakit diabetes. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%), nilai kuartil 2 (50%) atau median, dan nilai kuartil 3 (75%) memiliki nilai masing-masing 0 yang artinya *no* atau masyarakat tidak menderita penyakit diabetes. Sehingga variabel ini dapat dikatakan bahwa masyarakat secara umum masyarakat tidak menderita penyakit diabetes.

5.1.6.9 Analisis Deskripsi Data Variabel *Arthritis*

Tabel 5. 53 Analisis Deskripsi Data Variabel *Arthritis*

Variabel	*Count*	*Mean*	*Std*	*Min*	*25%*	*50%*	*75%*	*Max*
Arthritis	308.774	0.327304	0.469230	0.0	0.0	0.0	1.0	1.0

Berdasarkan tabel 5.53 Variabel *arthritis* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya tidak mengalami penyakit arthritis dan nilai maksimal atau terbesar dalam variabel yaitu 1 yang artinya mengalami penyakit arthritis. Nilai *mean* pada variabel ini yaitu 0.327304 dimana menggambarkan masyarakat tidak menderita penyakit arthritis. Nilai standar deviasi pada variabel ini yaitu 0.469230 yang menggambarkan pola persebaran data variabel ini. Nilai standar deviasi pada variabel memiliki nilai yang lebih besar dibandingkan dengan nilai *mean* yang menunjukan pola persebaran data yang lebih beragam dan nilai data tidak terlalu jauh dari nilai *mean*. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) dan nilai kuartil 2 (50%) atau median memiliki nilai masing-masing 0 yang artinya *no* atau masyarakat tidak menderita penyakit arthritis, sedangkan nilai kuartil 3 (75%) memiliki nilai masing-masing 1 yang artinya *yes* atau masyarakat menderita penyakit arthritis. Sehingga variabel ini dapat dikatakan bahwa masyarakat secara umum masyarakat tidak menderita penyakit arthritis.

5.1.6.10 Analisis Deskripsi Data Variabel *Sex*

Tabel 5. 54 Analisis Deskripsi Data Variabel *Sex*

Variabel	*Count*	*Mean*	*Std*	*Min*	*25%*	*50%*	*75%*	*Max*
Sex	308.774	0.481320	0.499652	0.0	0.0	0.0	1.0	1.0

Berdasarkan tabel 5.54 Variabel *sex* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya jenis kelamin perempuan dan nilai maksimal atau terbesar dalam variabel yaitu 1 yang artinya jenis kelamin laki-laki. Nilai *mean* pada variabel ini yaitu 0.481320 dimana menggambarkan masyarakat rata-rata memiliki jenis kelamin Perempuan. Nilai standar deviasi pada variabel ini yaitu 0.499652 yang menggambarkan pola persebaran data variabel ini. Nilai standar deviasi pada variabel memiliki nilai yang lebih besar dibandingkan dengan nilai *mean* yang menunjukan pola persebaran data yang lebih beragam dan nilai data tidak terlalu jauh dari

nilai *mean*. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) dan nilai kuartil 2 (50%) atau median memiliki nilai masing-masing 0 yang artinya *female* atau memiliki jenis kelamin perempuan, sedangkan nilai kuartil 3 (75%) memiliki nilai masing-masing 1 yang artinya *male* atau masyarakat memiliki jenis kelamin laki-laki. Sehingga variabel ini dapat dikatakan bahwa masyarakat secara umum masyarakat memiliki jenis kelamin perempuan.

5.1.6.11 Analisis Deskripsi Data Variabel *Age Category*

Tabel 5. 55 Analisis Deskripsi Data Variabel *Age Category*

Variabel	*Count*	*Mean*	*Std*	*Min*	*25%*	*50%*	*75%*	*Max*
Age_Category	308.774	6.536104	3.523495	0.0	4.0	7.0	9.0	12.0

Berdasarkan tabel 5.55 Variabel *age category* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 artinya rentang usia 18-24 tahun dan nilai maksimal atau terbesar dalam variabel yaitu 12 artinya retang usia 80 tahun keatas. Nilai *mean* pada variabel ini yaitu 6.536104 dimana menggambarkan rentang usia masyarakat pada variabel ini berada di posisi ditengah-tengah antara nilai minimal dan nilai maksimal. Nilai standar deviasi pada variabel ini yaitu 3.523495 yang menggambarkan pola persebaran data pada variabel ini. Nilai standar deviasi pada variabel ini dibawah nilai *mean*, sehingga persebaran dan keragaman data yang terdapat pada variabel ini kurang beragam dan tersebar. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) memiliki nilai 4 yang artinya memiliki rentang usia 40-44 tahun, nilai kuartil 2 (50%) atau median memiliki nilai 7 yang artinya memiliki rentang usia 50-59 tahun, dan nilai kuartil 3 (75%) yaitu 9 yang artinya memiliki rentang usia 65-69 tahun. Sehingga pada variabel ini dapat dikatakan bahwa rentang usia pada masyarakat secara umum yaitu 50-59 tahun.

5.1.6.12 Analisis Deskripsi Data Variabel Height (cm)

Tabel 5. 56 Analisis Deskripsi Data Variabel *Height (cm)*

Variabel	*Count*	*Mean*	*Std*	*Min*	*25%*	*50%*	*75%*	*Max*
Height_(cm)	308.774	170.615220	10.6584529	91.0	163.0	170.0	178.0	241.0

Berdasarkan tabel 5.56 Variabel *height (cm)* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 91 yang artinya tinggi badan

setinggi 91 cm dan nilai maksimal atau terbesar yaitu 241 yang artinya tinggi badan setinggi 241 cm. Nilai *mean* pada variabel ini yaitu 170.615220 dimana menggambarkan tinggi badan masyarakat secara umum dalam variabel ini yang posisinya berada ditengah-tengah antara nilai minimal dan nilai maksimal. Nilai standar deviasi pada variabel ini yaitu 10.6584529 yang menggambarkan pola persebaran data pada variabel ini. Nilai standar deviasi pada variabel ini dibawah nilai *mean*, sehingga persebaran dan keragaman data yang terdapat pada variabel ini kurang beragam dan tersebar. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) memiliki nilai 163 artinya tinggi badan masyarkat setinggi 163 cm, nilai kuartil 2 (50%) atau median memiliki nilai 170 artinya tinggi badan masyarkat setinggi 170 cm, nilai kuartil 3 (75%) yaitu 178 artinya tinggi badan masyarkat setinggi 178 cm. Sehingga pada variabel ini dapat dikatakan bahwa tinggi badan masyarakat rata-rata secara umum memiliki tinggi badan sekitar 170 cm.

5.1.6.13 Analisis Deskripsi Data Variabel *Weight (kg)*

Tabel 5. 57 Analisis Deskripsi Data Variabel *Weight (kg)*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
Weight_(kg)	308.774	83.590399	21.344664	24.95	68.04	81.65	95.25	293.02

Berdasarkan tabel 5.57 Variabel *weight (kg)* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 24.95 artinya berat badan seberat 24.95 kg dan nilai maksimal atau terbesar yaitu 293.02 yang artinya berat badan seberat 293.02 kg. Nilai *mean* pada variabel ini yaitu 83.590399 dimana menggambarkan berat badan masyarakat secara umum dalam variabel ini yang posisinya berada ditengah-tengah antara nilai minimal dan nilai maksimal. Nilai standar deviasi pada variabel ini yaitu 21.344664 yang menggambarkan pola persebaran data pada variabel ini. Nilai standar deviasi pada variabel ini dibawah nilai *mean*, sehingga persebaran dan keragaman data yang terdapat pada variabel ini kurang beragam dan tersebar. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) memiliki nilai 68.04 artinya berat badan masyarkat seberat 68.04 kg, nilai kuartil 2 (50%) atau median memiliki nilai 81.65 artinya berat badan masyarkat seberat 81.65 kg, nilai kuartil 3 (75%) yaitu 95.25 artinya berat badan masyarkat seberat 95.25 kg. Sehingga pada variabel ini dapat dikatakan bahwa tiberat badan masyarakat rata-rata secara umum memiliki berat badan sekitar 80 kg.

5.1.6.14 Analisis Deskripsi Data Variabel *BMI*

Tabel 5. 58 Analisis Deskripsi Data Variabel *BMI*

Variabel	*Count*	*Mean*	*Std*	*Min*	*25%*	*50%*	*75%*	*Max*
BMI	308.774	28.626813	6.522810	12.02	24.21	27.44	31.85	99.33

Berdasarkan tabel 5.58 Variabel *BMI* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 12.02 dan nilai maksimal atau terbesar dalam variabel yaitu 99.33. Nilai *mean* pada variabel ini yaitu 28.626813 dimana menggambarkan kondisi indeks masa tubuh masyarakat secara umum. Nilai standar deviasi pada variabel ini yaitu 6.522810 yang menggambarkan pola persebaran data pada variabel ini. Nilai standar deviasi pada variabel ini dibawah nilai *mean*, sehingga persebaran dan keragaman data yang terdapat pada variabel ini kurang beragam dan tersebar. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) memiliki nilai 24.21 yang artinya indeks masa tubuh masayarakat yaitu 24.21 Kg/m^2 , kemudian nilai kuartil 2 (50%) atau median memiliki nilai 27.44 yang artinya indeks masa tubuh masayarakat yaitu 27.44 Kg/m^2, dan nilai kuartil 3 (75%) yaitu 31.85 yang artinya indeks masa tubuh masayarakat yaitu 31.85 Kg/m^2 . Sehingga pada variabel ini dapat dikatakan bahwa indeks masa tubuh masyarakat secara umum berada di sekitar 28 Kg/m^2 .

5.1.6.15 Analisis Deskripsi Data Variabel *Smoking History*

Tabel 5. 59 Analisis Deskripsi Data Variabel *Smoking History*

Variabel	*Count*	*Mean*	*Std*	*Min*	*25%*	*50%*	*75%*	*Max*
Smoking_History	308.774	0.405662	0.491021	0.0	0.0	0.0	1.0	1.0

Berdasarkan tabel 5.59 Variabel *smoking history* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya masyarakat tidak merokok dan dan nilai maksimal atau terbesar dalam variabel yaitu 1 yang artinya masyarakat merokok. Nilai *mean* pada variabel ini yaitu 0.405662 dimana menggambarkan masyarakat rata-rata memiliki jenis kelamin Perempuan. Nilai standar deviasi pada variabel ini yaitu 0.491021yang menggambarkan pola persebaran data variabel ini. Nilai standar deviasi pada variabel memiliki nilai yang lebih besar dibandingkan dengan nilai *mean* yang menunjukan pola persebaran data yang lebih beragam dan nilai data tidak terlalu jauh dari nilai *mean*. Nilai kuartil pada variabel ini

terdiri dari kuartil 1 (25%) dan nilai kuartil 2 (50%) atau median memiliki nilai masing-masing 0 yang artinya jenis kelamin masyarakat perempuan, sedangkan nilai kuartil 3 (75%) memiliki nilai 1 yang artinya jenis kelamin laki-laki. Sehingga variabel ini dapat dikatakan secara umum jenis kelamin yang banyak di Masyarakat yaitu Perempuan.

5.1.6.16 Analisis Deskripsi Data Variabel *Alcohol Consumption*

Tabel 5. 60 Analisis Deskripsi Data Variabel *Alcohol Consumption*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
Alcohol_Consumption	308.774	5.097557	8.200434	0.0	0.0	1.0	6.0	30.0

Berdasarkan tabel 5.60 Variabel *alcohol consumption* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya konsumsi alkohol dengan frekuensi sebanyak 0 kali dan nilai maksimal atau terbesar dalam variabel yaitu 30 yang artinya konsumsi alkohol dengan frekuensi sebanyak 30 kali. Nilai *mean* pada variabel ini yaitu 5.097557 dimana menggambarkan konsumsi alkohol masyarakat secara umum dengan frekuensi sekitar 5 kali. Nilai standar deviasi pada variabel ini yaitu 8.200434 yang menggambarkan pola persebaran data variabel ini. Nilai standar deviasi pada variabel memiliki nilai yang lebih besar dibandingkan dengan nilai *mean* yang menunjukan pola persebaran data yang lebih beragam. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) sebesar 0 yang artinya 0 kali, nilai kuartil 2 (50%) atau median sebesar 1 yang artinya 1 kali, dan nilai kuartil 3 (75%) memiliki nilai sebesar 6 yang artinya 6 kali. Sehingga variabel ini dapat dikatakan bahwa masyarakat secara umum konsumsi alkohol masyarakat dengan jumlah frekuensi yang rendah.

5.1.6.17 Analisis Deskripsi Data Variabel *Fruit Consumption*

Tabel 5. 61 Analisis Deskripsi Data Variabel *Fruit Consumption*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
Fruit_Consumption	308.774	29.834290	24.877812	0.0	12.0	30.0	30.0	120.0

Berdasarkan tabel 5.61 Variabel *fruit consumption* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya konsumsi buah dengan frekuensi sebanyak 0 kali dan nilai maksimal atau terbesar dalam variabel yaitu 120 yang artinya konsumsi buah dengan frekuensi sebanyak 120 kali. Nilai *mean* pada variabel ini yaitu 29.834290 dimana menggambarkan konsumsi buah oleh

masyarakat. Nilai standar deviasi pada variabel ini yaitu 24.877812 yang menggambarkan pola persebaran data pada variabel ini. Nilai standar deviasi pada variabel ini dibawah nilai *mean*, sehingga persebaran dan keragaman data yang terdapat pada variabel ini kurang beragam dan tersebar. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) memiliki nilai 12 yang artinya konsumsi buah masyarakat sebanyak 12 kali, nilai kuartil 2 (50%) atau median dan nilai kuartil 3 (75%) yaitu 30 yang artinya konsumsi buah sebanyak 30 kali. Sehingga pada variabel ini dapat dikatakan bahwa frekuensi konsumsi buah di masyarakat berada di posisi tengah-tengah dari total masyarakat.

5.1.6.18 Analisis Deskripsi Data Variabel *Green Vegetables Consumption*

Tabel 5. 62 Analisis Deskripsi Data Variabel *Green Vegetables Consumption*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
Green_Vegetables_Consumption	308.774	15.109517	14.926912	0.0	4.0	12.0	20.0	128.0

Berdasarkan tabel 5.62 Variabel *green vegetables consumption* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya jumlah konsumsi sayur hijau dengan frekuensi sebanyak 0 kali dan nilai maksimal atau terbesar dalam variabel yaitu 128 artinya jumlah konsumsi sayur hijau dengan frekuensi sebanyak 128 kali. Nilai *mean* pada variabel ini yaitu 15.109517 dimana menggambarkan konsumsi sayuran hijau oleh masyarakat. Nilai standar deviasi pada variabel ini yaitu 14.926912 yang menggambarkan pola persebaran data pada variabel ini. Nilai standar deviasi pada variabel ini dibawah nilai *mean*, sehingga persebaran dan keragaman data yang terdapat pada variabel ini kurang beragam dan tersebar. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) memiliki nilai 4 yang artinya konsumsi sayuran hijau oleh masyarakat sebanyak 4 kali, nilai kuartil 2 (50%) atau median memiliki nilai 12 yang artinya konsumsi sayuran hijau oleh masyarakat sebanyak 12 kali, dan nilai kuartil 3 (75%) yaitu 20 yang artinya konsumsi sayuran hijau oleh masyarakat sebanyak 20 kali. Sehingga pada variabel ini dapat dikatakan bahwa frekuensi konsumsi sayuran hijau di masyarakat berada di posisi tengah-tengah dari total masyarakat.

5.1.6.19 Analisis Deskripsi Data Variabel *Freid Potato Consumption*

Tabel 5. 63 Analisis Deskripsi Data Variabel *Fried Potato Consumption*

Variabel	***Count***	***Mean***	***Std***	***Min***	***25%***	***50%***	***75%***	***Max***
FriedPotato_ Consumption	308.774	6.297237	8.583837	0.0	2.0	4.0	8.0	128.0

Berdasarkan tabel 5.63 Variabel *freid potato consumption* terdiri dari 308.774 baris yang didalamnya terdapat nilai minimal atau terkecil dalam variabel yaitu 0 yang artinya jumlah konsumsi kentang goreng dengan frekuensi sebanyak 0 kali dan nilai maksimal atau terbesar dalam variabel yaitu 128 artinya jumlah konsumsi kentang goreng dengan frekuensi sebanyak 128 kali. Nilai *mean* pada variabel ini yaitu 6.297237 dimana menggambarkan konsumsi kentang goreng oleh masyarakat. Nilai standar deviasi pada variabel ini yaitu 8.583837 yang menggambarkan pola persebaran data pada variabel ini. Nilai standar deviasi pada variabel ini dibawah nilai *mean*, sehingga persebaran dan keragaman data yang terdapat pada variabel ini kurang beragam dan tersebar. Nilai kuartil pada variabel ini terdiri dari kuartil 1 (25%) memiliki nilai 2 yang artinya konsumsi kentang goreng oleh masyarakat sebanyak 2 kali, nilai kuartil 2 (50%) atau median memiliki nilai 4 yang artinya konsumsi kentang goreng oleh masyarakat sebanyak 4 kali, dan nilai kuartil 3 (75%) yaitu 8 yang artinya konsumsi kentang goreng oleh masyarakat sebanyak 8 kali. Sehingga pada variabel ini dapat dikatakan bahwa frekuensi konsumsi kentang goreng oleh masyarakat secara umum cenderung rendah.

5.1.7 Analisis Hasil *Data Training* dan *Data Testing*

Data training dilakukan untuk memberikan pelatihan bagi data dalam mempelajari berbagai pola data maupun mengoptimalkan data yang digunakan dalam pemodelan *AutoGluon*. Sedangkan *testing data* digunakan untuk menilai evaluasi kinerja model yang sebelumnya telah dilatih. *Training data* dan *testing data* dilakukan *split data* atau pemisahan data antara data yang digunakan untuk training maupun testing yang nantinya dilakukan pembobotan dengan perbandingan rasio 7:3. Perbandingan rasio 7:3 artinya 70% pembobotan untuk data *training* dan 30% pembobotan untuk data *testing*. Penelitian ini menggunakan perbandingan rasio 7:3 dibandingkan dengan menggunakan 8:2 atau

80% untuk data *training* dan 20% untuk data *testing* karena nilai akurasi yang dihasilkan dengan perbandingan rasio 7:3 memiliki nilai yang lebih tinggi dibandingkan dengan perbandingan rasio 8:2. Hal ini telah dilakukan oleh (Muhammad Raffi, 2023) dengan judul penelitian Analisis Sentimen Ulasan Aplikasi Binar Pada *Google Play Store* Menggunakan Algoritma *Naive Bayes* dimana hasil pengujian dengan perbandingan rasio 7:3 mendapatkan nilai akurasi sebesar 91,12%, sedangkan pengujian dengan perbandingan rasio 8:2 mendapatkan nilai akurasi sebesar 89,51%. Selain itu juga penelitian yang dilakukan oleh (Jung Hyun An, 2023) dengan judul *A CNN-Based Automatic Vulnerability* dalam hasil pengujian dengan perbandingan 7:3 memperoleh hasil nilai akurasi sebesar 98%, sedangkan perbandingan 8:2 memperoleh nilai akurasi sebesar 97%. Dalam mengatasi nilai yang mengalami ketidaknormalan dilakukan penerapan dengan metode *SMOTE*. Sehingga dengan penerapan *SMOTE* diharapkan dapat mengatasi ketidaknormalan dalam ketidakseimbangan kelas dalam *dataset*. Berikut ini merupakan hasil dari *data training* dan *data testing*:

```
Jumlah data trainig sebanyak: 216141
Jumlah data testing sebanyak: 92633
```

Gambar 5. 78 Hasil *Data Training* dan *Data Testing* tanpa SMOTE

```
Jumlah data trainig sebanyak: 397324
Jumlah data testing sebanyak: 170282
```

Gambar 5. 79 Hasil *Data Training* dan *Data Testing* dengan SMOTE

Berdasarkan gambar diatas, dapat diketahui jumlah *data training* lebih banyak dibandingkan *data testing* dengan nilai *data training* sebanyak 397.324 data dan nilai *data test* sebanyak 170.282 data. Nilai yang berbeda dikarenakan adanya contoh kelas pada minoritas ditambahkan dengan sintetis dalam mengatasi ketidakseimbangan data pada kelas minoritas yang memiliki jumlah yang sedikit. Kemudian pemisahan data menjadi *data training* dan *data testing* pada kolom *Heart_Disease* menggunakan perbandingan 7:3. Nilai perbandingan 7:3 menunjukan bahwa jumlah *data training* yang digunakan sebanyak 70% dari data sebenarnya dan *data testing* sebanyak 30% dari data sebenarnya. Penggunaan nilai perbandingan 7:3 bertujuan untuk mengetahui kemampuan *machine learning* dalam melakukan prediksi akurasi data yang lebih baik.

5.2 Analisis Hasil Pemodelan *AutoGluon*

5.2.1 Hasil Prediksi Pemodelan *AutoGluon*

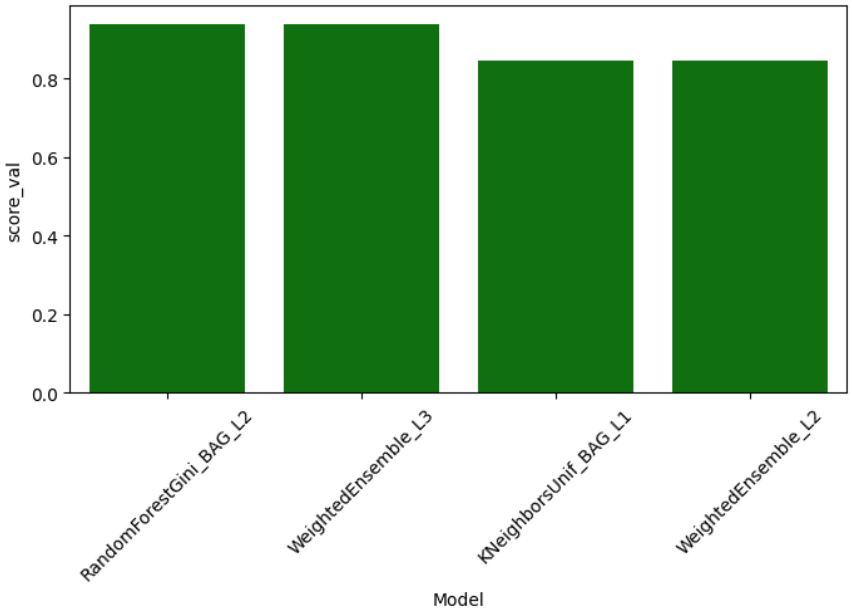
Model *AutoGluon* dalam penelitian ini menggunakan label *Heart_Disease* sebagai target untuk melakukan prediksi. Prediksi yang dilakukan dengan jenis *supervised learning* untuk melakukan klasifikasi dalam mencari nilai akurasi. Selain itu juga prediksi yang dilakukan untuk mencari metode terbaik dalam melakukan prediksi. Berikut hasil pemodelan dengan menggunakan *AutoGluon*:

	model	score_val	eval_metric	pred_time_val	fit_time	pred_time_val_marginal	fit_time_marginal	stack_level	can_infer	fit_order
0	RandomForestGini_BAG_L2	0.939724	accuracy	88.657702	56.108116	2.667743	52.559307	2	True	3
1	WeightedEnsemble_L3	0.939724	accuracy	89.653601	77.460698	0.995899	21.352582	3	True	4
2	KNeighborsUnif_BAG_L1	0.845119	accuracy	85.989959	3.548810	85.989959	3.548810	1	True	1
3	WeightedEnsemble_L2	0.845119	accuracy	86.627817	3.641543	0.637858	0.092733	2	True	2

Gambar 5. 80 Hasil Pemodelan *AutoGluon*

```
WeightedEnsemble_L3
```

Gambar 5. 81 Model *AutoGluon* Terbaik



Gambar 5. 82 Perbandingan nilai valensi berdasarkan model

Berdasarkan Gambar 5.80, terdapat beberapa model yang dihasilkan oleh model *AutoGluon* pada label *Heart_Disease* yang terdiri dari *RandomForestGini_BAG_L2*, *WeightedEnsemble_L3*, *KNeigborsUnif_BAG_L1*, *WeightedEnsemble_L2*. *Ensemble learning* sendiri merupakan penggabungan dari beberapa metode berdasarkan hasil dalam model *machine learning*. Penggunaan *ensemble learning* dapat memberikan keuntungan

dalam meningkatkan hasil akurasi maupun mengurangi *testing* secara berlebihan (Rifqi Kurniawan, 2015). *Weighted ensemble learning* melakukan penyesuaian dengan menyatukan hasil prediksi pada model yang telah dilatih sebelumnya (AutoGluon 1.0.0 Documentation, 2023).

Penelitian ini dilakukan dengan *prest best_quality* yang bertujuan untuk mencari model *machine learning* pada *AutoGluon* dengan hasil terbaik. Model terbaik berdasarkan penerapan *AutoGluon* berdasarkan pada penelitian ini yaitu *WeightedEnsemble_L3* dan *RandomForestGini_BAG_L2* yang masing-masing memiliki nilai *score_val* 0.939724 dimana hasil ini memiliki pengaruh dalam waktu prediksi maupun waktu pelatihan. Nilai *score_val* menggambarkan nilai validasi kinerja model *AutoGluon* berdasarkan *eval_metric* yang digunakan dalam penelitian ini yaitu *accuracy*. Sehingga nilai *score_val* yang dihasilkan memiliki jumlah nilai yang tidak berbeda jauh dengan nilai akurasi. Walaupun pada model *WeightedEnsemble_L3* dan *RandomForestGini_BAG_L2* memiliki nilai *score_val* yang sama, tetapi untuk model terbaik terdapat pada model *WeightedEnsemble_L3*.

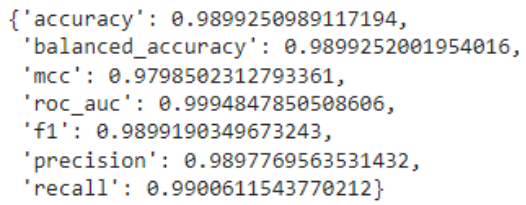
Pemilihan *WeightedEnsemble_L3* sebagai model terbaik dapat dilihat melalui *Pred_time_val*, dimana *Pred_time_val* menggambarkan waktu yang dibutuhkan dalam melakukan prediksi dengan nilai validasi pada model *WeightedEnsemble_L3* membutuhkan waktu selama 89.653601 sekon sedangkan model *RandomForestGini_BAG_L2* membutuhkan waktu selama 88.657702. Melalui waktu prediksi paling lama dalam melakukan validasi dibandingkan dengan model *RandomForestGini_BAG_L2*, sehingga model *WeightedEnsemble_L3* dianggap kurang efisien.

Fit_time dalam penelitian ini pada model *RandomForestGini_BAG_L2* membutuhkan waktu sekitar 56.108116 sekon, sedangkan pada model *WeightedEnsemble_L3* membutuhkan waktu sekitar 77.460698 sekon. Hal ini menggambarkan waktu yang dibutuhkan dalam melatih model untuk meningkatkan kinerja model. Nilai *score_val* pada model *WeightedEnsemble_L3* memiliki nilai yang paling tinggi dibandingkan dengan model yang lain, sehingga waktu yang dibutuhkan dalam *fit_time* juga paling lama.

Pred_time_val_marginal dalam model *WeightedEnsemble_L3* membutuhkan waktu selama 0.995899 sekon, sedangkan model *RandomForestGini_BAG_L2* membutuhkan waktu selama 2.66743 sekon. Nilai *Pred_time_val_marginal* menggambarkan informasi waktu tambahan dalam menghitung prediksi dengan nilai validasi. Melalui waktu tambahan yang rendah menandakan bahwa model *WeightedEnsemble_L3* dapat berjalan dengan baik.

Fit_time_marginal memberikan gambaran mengenai informasi waktu yang dibutuhkan dalam penyesuaian durasi waktu dalam melatih model. Model *WeightedEnsemble_L3* membuhtukan waktu selama 21.352582 sekon dan model *RandomForestGini_BAG_L2* membutuhkan waktu selama 52.559307 sekon. Sehingga model *WeightedEnsemble_L3* membutuhkan waktu dalam melatih model yang lebih singkat.

Stack_level pada model *WeightedEnsemble_L3* mendapatkan nilai sebesar 3 dan pada model *RandomForestGini_BAG_L2* mendapatkan nilai 2. Nilai *stack_level* pada model *WeightedEnsemble_L3* menggambarkan tingkat kompleksitas pada *ensemble learning* yang lebih baik dibandingkan dengan model *RandomForestGini_BAG_L2*. *Can_infer* pada penelitian ini mendapatkan nilai *True,* sehingga model penelitian ini dapat melakukan inferensi data. *Fit_order* yang pada model *WeightedEnsemble_L3* mendapatkan nilai sebesar 4 dan pada model *RandomForestGini_BAG_L2* mendapatkan nilai 3. Nilai *Fit_oder* ini menandakan urutan model yang dilatih pada model *WeightedEnsemble_L3* berada di urutan 4 dan model *RandomForestGini_BAG_L2* berada di urutan 3.



Gambar 5. 83 Nilai hasil *evaluasi AutoGluon*

Berdasarkan Gambar 5.83, dapat diketahui hasil evaluasi yang telah dilakukan pada hasil model pengujian kali ini. Hasil evaluasi yang dilakukan dalam *AutoGluon* dilakukan untuk memahami *eval_metric* yang digunakan. Evaluasi dalam penelitian ini dilakukan

dengan *eval_metric* untuk melakukan validasi pada model yang dihasilkan (AutoGluon 1.0.0 Documentation, 2023). Nilai *accuracy* mendapatkan nilai sebesar 0.9899250989117194, nilai ini menggambarkan kinerja model dalam melakukan prediksi dimana dalam penelitian ini nilai *accuracy* dan nilai *score_val* memiliki nilai perbedaan yang tidak jauh. Kemudian dapat diketahui nilai *balance_accuracy* mendapatkan nilai sebesar 0.9899252001954016, nilai ini menggambarkan keseimbangan pada nilai akurasi. Keseimbangan ini dapat dilihat pada nilai *accuracy* dan nilai *balanced_accuracy* memiliki yang tidak berebeda jauh. Selanjutnya terdapat nilai *mcc (Matthews Correlation Coeficient)* mendapatkan nilai sebesar 0.9798502312793361, nilai ini menggambarkan kinerja pada algoritma klasifikasi ketika melakukan prediksi. Lalu terdapat nilai *roc_auc (Receiver Operating Characteristic – Area Under Curve)* mendapatkan nilai 0.9994847850508606, nilai *roc_auc* dalam penelitian ini menggambarkan probabilitas pada kinerja model. Selanjutnya terdapat nilai *f1* mendapatkan nilai sebesar 0.9899190349673243, nilai *f1* atau *f1-score* yang didapatkan menggambarkan nilai rata-rata antara nilai *precision* dan nilai *recall*. Hal ini dapat dilihat melalui perbedaan nilai *recall* dan nilai *precision* yang tidak berbeda jauh. Nilai *precision* dalam penelitian ini mendapatkan nilai sebesar 0.9897769563531432, nilai ini memberikan illustasi mengenai kesesuaian nilai prediksi yang benar dan nilai positif yang dihasilkan. Nilai *recall* dalam penelitian ini mendapatkan nilai sebesar 0.9900611543770212, nilai ini memberikan illustasi mengenai kemampuan model dalam melakukan deteksi dengan nilai yang benar dan positif.

5.2.2 Perbandingan Nilai Akurasi Model

Berdasarkan prediksi yang dilakukan dengan pemodelan *AutoGluon*, maka dapat diketahui hasil model yang dihasilkan yaitu *RandomForestGini_BAG_L2*, *WeightedEnsemble_L3, KNeigborsUnif_BAG_L1, WeightedEnsemble_L2*. Berikut merupakan perbandingan nilai akurasi dalam menentukan model klasifikasi terbaik.

```
Nilai Akurasi  Random Forest: 0.93
Hasil Klasifikasi Random Forest:
              precision    recall  f1-score   support

           0       0.93      0.93      0.93     84993
           1       0.93      0.93      0.93     85289

    accuracy                           0.93    170282
   macro avg       0.93      0.93      0.93    170282
weighted avg       0.93      0.93      0.93    170282
```

Gambar 5. 84 Nilai Akurasi model *Random Forest*

```
Nilai Akurasi K-NN: 0.87
Hasil Klasifikasi k-NN:
              precision    recall  f1-score   support

           0       0.99      0.74      0.85     84993
           1       0.79      1.00      0.88     85289

    accuracy                           0.87    170282
   macro avg       0.89      0.87      0.87    170282
weighted avg       0.89      0.87      0.87    170282
```

Gambar 5. 85 Nilai Akurasi model *K-NN*

```
Nilai Akurasi Weighted Ensemble Learning: 0.94
Hasil Klasifikasi Weighted Ensemble Learning:
              precision    recall  f1-score   support

           0       0.97      0.91      0.94     84993
           1       0.91      0.97      0.94     85289

    accuracy                           0.94    170282
   macro avg       0.94      0.94      0.94    170282
weighted avg       0.94      0.94      0.94    170282
```

Gambar 5. 86 Nilai Akurasi model *Weighted Ensemble Learning*

Berdasarkan hasil prediksi yang telah dilakukan maka dapat diketahui bahwa nilai akurasi tertinggi terdapat pada model *Weighted Ensemble Learning* dengan nilai 0.94, kemudian pada model *Random Forest* dengan nilai 0.93, dan model *K-NN* dengan nilai 0.87. *Weighted Ensemble Learning* mendapatkan nilai akurasi tertinggi karena adanya integrasi hasil dari model *machine learning* yang berbeda-beda yang memberikan bobot berlebih pada model *Weighted Ensemble Learning*. Integrasi pada model *Weighted Ensemble Learning* menandakan model ini dapat memaksimalkan dalam mengoptimalkan performa model. Selain itu juga nilai akurasi yang tinggi pada model *Weighted Ensemble Learning* menandakan model memperoleh nilai yang akurat. Kemudian nilai akurasi model *Random Forest* termasuk dalam model yang kuat karena memiliki nilai tidak berbeda jauh berbeda dengan model *Weighted Ensemble Learning*. Selanjutnya model *K-NN* memiliki nilai akurasi yang rendah menandakan model tersebut kurang efektif dalam

memprediksi pola data tertentu dibandingkan dengan model *Weighted Ensemble Learning* dan model *Random Forest*.

5.2.3 Hasil Prediksi Penyakit Jantung

Berikut merupakan hasil prediksi jantung dengan menggunakan *machine learning autogluon*:

1. Kondisi kesehatan masyarakat mayoritas dalam keadaan baik dan tidak terkena penyakit jantung. Tetapi kondisi kesehatan yang buruk dan terkena penyakit jantung menjadi perhatian khusus karena rentan terhadap terkena penyakit.
2. Mayoritas masyarkat melakukan pemeriksaan rutin pada setiap tahun sehingga dapat mengetahui kondisi yang sehat dan tidak terkena penyakit jantung. Tetapi dengan melakukan pemeriksaan rutin setiap tahun membantu dalam mengetahui jumlah dan persentase penyakit jantung yang tinggi.
3. Latihan fisik yang telah dilakukan secara mayoritas masyarakat tidak terkena penyakit jantung walaupun juga dengan melakukan latihan fisik masih terdapat risiko penyakit jantung. Namun masyarakat yang tidak melakukan latihan fisik lebih rentan terkena penyakit jantung.
4. Mayoritas masyarakat tidak menderita penyakit kanker kulit maupun tidak menderita penyakit jantung. Namun masyarakat yang menderita penyakit kanker kulit menjadi perhatian khusus karena rentan terkena penyakit jantun.
5. Masyarakat tidak menderita penyakit kanker dengan jenis yang lain maupun penyakit jantung secara mayoritas. Tetapi masyarakat yang memiliki penyakit kanker dengan jenis yang lain lebih rentan terkena penyakit jantung.
6. Masyarakat yang tidak depresi secara umum tidak menderita penyakit jantung walaupun juga masih terdapat risiko terkena penyakit jantung. Akan tetapi masyarakat yang menderita depresi lebih rentan terkena penyakit jantung yang menjadi perhatian khusus.
7. Mayoritas masyarakat tidak terkena penyakit diabetes maupun tidak terkena penyakit jantung. Namun masyarakat yang terkena penyakit diabetes lebih rentan terkena penyakit jantung.

8. Masyarakat secara mayoritas tidak terkena penyakit arthritis maupun tidak terkena penyakit jantung. Tetapi masyarakat yang mengalami penyakit arthritis lebih rawan terkena penyakit jantung.
9. Jenis kelamin perempuan secara mayoritas tidak terkena penyakit jatung walaupun masih terdapat risiko terkena penyakit jantung. Akan tetapi jenis kelamin laki-laki lebih rentan terkena penyakit jantung dengan persentase yang lebih tinggi dibandingkan Perempuan.
10. Secara mayoritas masyarakat yang tidak merokok tidak terkena penyakit jantung walaupun terdapat kemungkinan terkena penyakit jantung. Tetapi masyarakat yang merokok lebih rawan terkena penyakit jantung yang menjadi perhatian khusus.
11. Kategori usia masyrakat yang rentan terkena penyakit jantung berada pada usia diatas 80 tahun keatas. Kemudian tinggi dan berat badan masyarakat mayoritas memiliki tinggi 168 cm dan memiliki berat badan 90.72 kg. BMI (*Body Mass Index*) pada masyarakat secara mayoritas memiliki nilai 26.63 Kg/m^2.
12. Konsumsi alkohol di masyarakat memiliki jumlah frekuensi sebanyak 0 kali yang menandakan masyarakat secara umum tidak mengkonsumsi alkohol. Konsumsi buah dan sayuran hijau di masyarakat memiliki jumlah frekuensi sebanyak 30 kali yang menandakan konsumsi buah dan sayuran hijau hampir dikonsumsi setiap oleh masyarakat dalam satu bulan. Konsumsi kentang goreng di masyarakat memiliki jumlah frekuensi sebanyak 4 kali yang menandakan dalam sebulan paling tidak dalam satu minggu mengkonsumsi kentang goreng sebanyak 1 kali.

5.3 Analisis Hasil Evaluasi Model

5.3.1 *Confusion Matrix*

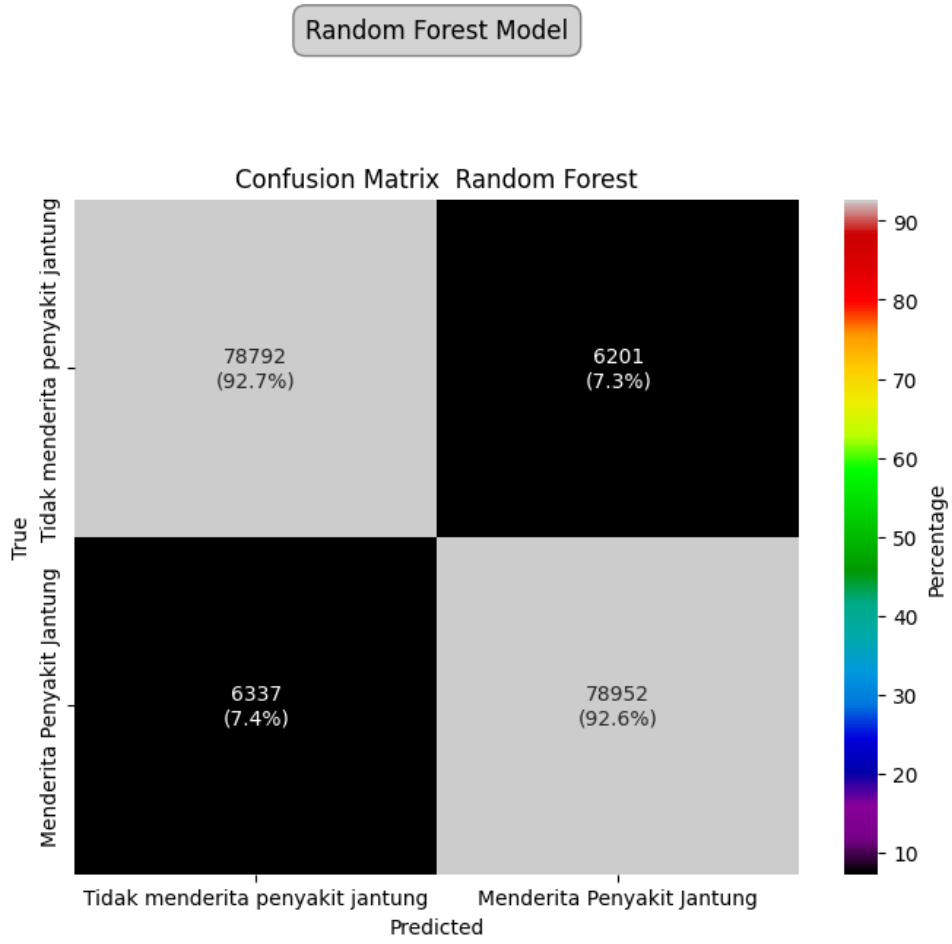
Evaluasi model yang dilakukan dengan *confusion matrix* bertujuan untuk melakukan evaluasi kinerja model yang telah dilakukan dalam menguji hasil klasifikasi prediksi. Perancangan *confusion matrix* dilakukan dengan normalisasi dengan menggunakan persentase untuk melakukan koversi dari nilai numerik menjadi nilai persentase.

5.3.1.1 *Confusion Matrix* model *Random Forest*

Berikut merupakan hasil evaluasi model dengan menggunakan *confusion matrix* pada model *Random Forest*:

Tabel 5. 64 Evaluasi dengan *Confusion Matrix* model *Random Forest*

Prediksi	**Positive**	**Negative**
Positive	True Positive (TP)	False Positive (FP)
	78.792	6.201
	92.7%	7.3%
Negative	False Negative (FN)	True Negative (TN)
	6.337	78.952
	7.4%	92.6%



Gambar 5. 87 Hasil *Confusion Matrix* model *Random Forest*

Berdasarkan Gambar 5.87 terdapat keterangan *True Positive* (TP) yang menandakan jumlah masyarakat yang dinyatakan tidak menderita penyakit jantung dan tidak menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan tidak menderita penyakit jantung dan benar tidak menderita penyakit jantung sebanyak 78.792 orang dengan persentase 92.7%.

Kemudian terdapat *False Positive* (FP) yang menandakan jumlah masyarakat yang dinyatakan menderita penyakit jantung tetapi kebenarannya tidak menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan menderita penyakit jantung dan kebenarannya tidak menderita penyakit jantung sebanyak 6.201 orang dengan persentase 7.3%.

Selanjutnya terdapat *True Negative* (TN) yang menandakan jumlah masyarakat yang dinyatakan menderita penyakit jantung dan benar-benar menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan menderita penyakit jantung dan benar menderita penyakit jantung sebanyak 78.952 orang dengan persentase 92.6%.

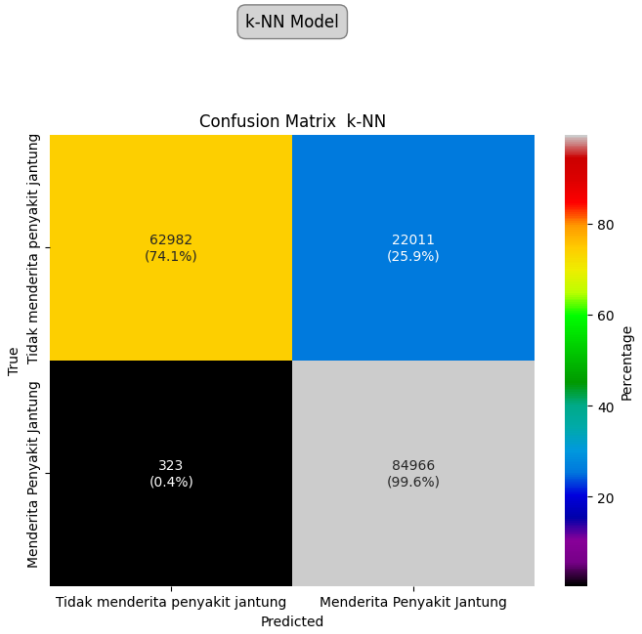
Lalu terdapat *False Negative* (FN) yang menandakan jumlah masyarakat yang dinyatakan tidak menderita penyakit jantung tetapi kebenarannya menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan mendertia penyakit jantung dan tidak benar menderita penyakit jantung sebanyak 36.337orang dengan persentase 7.4%.

5.3.1.2 *Confusion Matrix* model *K-NN*

Berikut merupakan hasil evaluasi model dengan menggunakan *confusion matrix* pada model *K-NN*:

Tabel 5. 65 Evaluasi dengan *Confusion Matrix* model *K-NN*

Prediksi	**Positive**	**Negative**
Positive	True Positive (TP)	False Positive (FP)
	62.982	22.011
	74.1%	25.9%
Negative	False Negative (FN)	True Negative (TN)
	323	84.966
	0.4%	99.6%



Gambar 5. 88 Hasil *Confusion Matrix* model *K-NN*

Berdasarkan Gambar 5.88 terdapat keterangan *True Positive* (TP) yang menandakan jumlah masyarakat yang dinyatakan tidak menderita penyakit jantung dan tidak menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan tidak menderita penyakit jantung dan benar tidak menderita penyakit jantung sebanyak 62.982 orang dengan persentase 74.1%.

Kemudian terdapat *False Positive* (FP) yang menandakan jumlah masyarakat yang dinyatakan menderita penyakit jantung tetapi kebenarannya tidak menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan menderita penyakit jantung dan kebenarannya tidak menderita penyakit jantung sebanyak 22.011 orang dengan persentase 25.9%.

Selanjutnya terdapat *True Negative* (TN) yang menandakan jumlah masyarakat yang dinayatakan menderita penyakit jantung dan benar-benar menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan menderita penyakit jantung dan benar menderita penyakit jantung sebanyak 84.966 orang dengan persentase 99.6%.

Lalu terdapat *False Negative* (FN) yang menandakan jumlah masyarakat yang dinyatakan tidak menderita penyakit jantung tetapi kebenarannya menderita penyakit

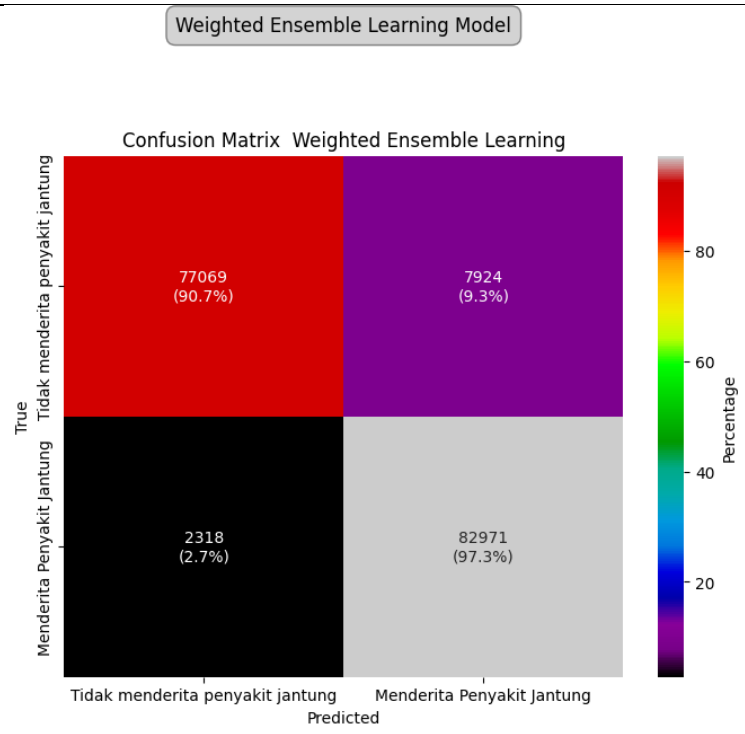
jantung. Dalam hal ini masyarakat yang dinyatakan mendertia penyakit jantung dan tidak benar menderita penyakit jantung sebanyak 323 orang dengan persentase 0.4%.

5.3.1.3 *Confusion Matrix* model *Weighted Ensemble Learning*

Berikut merupakan hasil evaluasi model dengan menggunakan *confusion matrix* pada model *Weighted Ensemble Learning*:

Tabel 5. 66 Evaluasi dengan *Confusion Matrix* model *Weighted Ensemble Learning*

Prediksi	**Positive**	**Negative**
Positive	True Positive (TP)	False Positive (FP)
	77.069	7.924
	90.7%	9.3%
Negative	False Negative (FN)	True Negative (TN)
	2.318	82.971
	2.7%	97.3%



Gambar 5. 89 Hasil *Confusion Matrix* model *Weighted Ensemble Learning*

Berdasarkan Gambar 5.89 terdapat keterangan *True Positive* (TP) yang menandakan jumlah masyarakat yang dinyatakan tidak menderita penyakit jantung dan tidak menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan tidak menderita penyakit

jantung dan benar tidak menderita penyakit jantung sebanyak 77.069 orang dengan persentase 90.7%.

Kemudian terdapat *False Positive* (FP) yang menandakan jumlah masyarakat yang dinyatakan menderita penyakit jantung tetapi kebenarannya tidak menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan menderita penyakit jantung dan kebenarannya tidak menderita penyakit jantung sebanyak 7.924 orang dengan persentase 9.3%.

Selanjutnya terdapat *True Negative* (TN) yang menandakan jumlah masyarakat yang dinayatakan menderita penyakit jantung dan benar-benar menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan menderita penyakit jantung dan benar menderita penyakit jantung sebanyak 82.971 orang dengan persentase 97.3%.

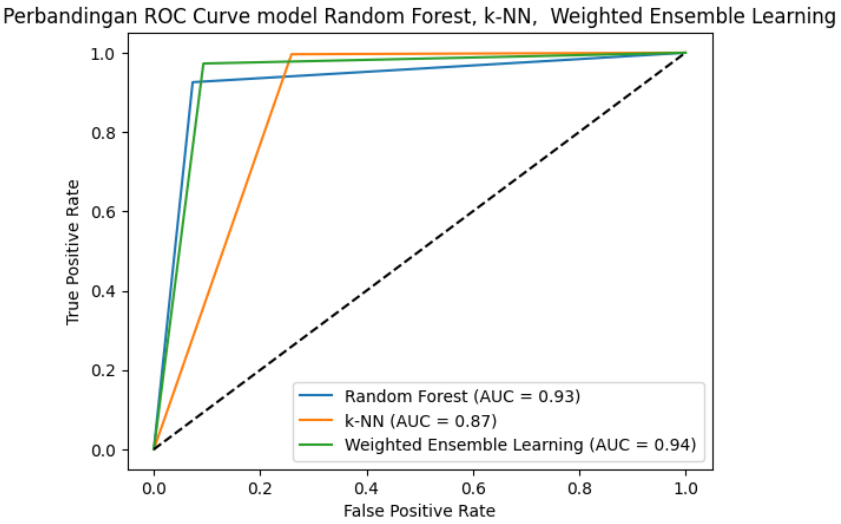
Lalu terdapat *False Negative* (FN) yang menandakan jumlah masyarakat yang dinyatakan tidak menderita penyakit jantung tetapi kebenarannya menderita penyakit jantung. Dalam hal ini masyarakat yang dinyatakan mendertia penyakit jantung dan tidak benar menderita penyakit jantung sebanyak 2.318 orang dengan persentase 2.7%

5.3.1.4 Hasil *Confusion Matrix*

Berdasarkan pengujian evaluasi model yang dilakukan, dapat diketahui pada model *random forest, k-nn,* dan *weighted ensemble learning* memiliki kelebihan dan kekurangan masing-masing. Akan tetapi jika dilihat dari nilai akurasi, maka model *weighted ensemble learning* memiliki keakuratan yang paling tinggi karena memiliki nilai akurasi yang paling tinggi dibanding dengan model yang lainnya.

5.3.2 *ROC Curve*

Evaluasi model yang dilakukan dengan menggunakan *ROC Curve* bertujuan untuk tingkat sensitivitas pada model. Selain itu juga untuk mengetahui kinerja model klasifikasi dengan optimal. Berikut merupakan hasil evaluasi model dengan menggunakan *ROC Curve*.



Gambar 5. 90 *ROC Curve*

Berdasarkan Gambar 5.90, dapat diketahui bahwa model yang memiliki nilai *AUC* (*Area Under Curve*) dengan nilai tertinggi yaitu model *Weighted Ensemble Learning* dengan nilai 0.94, kemudian model *Random Forest* dengan nilai 0.93, dan model *K-NN* dengan nilai 0.87. Model *Weighted Ensemble Learning* menjadi model terbaik karena memiliki nilai mendekati 1 yang menandakan kinerja pada model ini telah berjalan dengan baik. Nilai *AUC* yang tinggi menandakan model dapat menjalankan klasifikasi kelas positif dan kelas negatif dengan baik. Nilai *Weighted Ensemble Learning* menjadi nilai terbaik karena menggabungkan hasil prediksi dari beberapa model. Nilai *AUC Random Forest* yaitu 0.93 yang termasuk dalam model kuat dengan dapat menjalankan klasifikasi kelas positif dan kelas negatif dengan baik walaupun sedikit dibawah model *Weighted Ensemble Learning*. Kemudian nilai *AUC* termasuk kedalam nilai yang kurang baik dalam menjalan klasifikasi kelas positif dan kelas negatif dibandingkan dengan model yang lainnya karena memiliki niali *AUC* terendah yaitu 0,87.

BAB VI

PENUTUP

6.1 Kesimpulan

Berdasarkan informasi yang dikumpulkan, berikut merupakan kesimpulan untuk menjawab rumusan masalah sebagai berikut:

1. Hasil prediksi penyakit jantung berdasarkan pada dataset *cardiovascular risk disease prediction dataset* yang dilakukan dengan nilai perbandingan antara *data training* dan *data testing* sebesar 7:3 diperoleh model yaitu model *RandomForestGini_BAG_L2*, *WeightedEnsemble_L3, KNeigborsUnif_BAG_L1,* dan *WeightedEnsemble_L2.*
2. Model terbaik berdasarkan pemodelan *AutoGluon* yang paling akurat dilakukan dengan target label *Heart_Disease* dan nilai perbandingan antara *data training* dan *data testing* sebesar 7:3 yaitu terdapat pada model *WeightedEnsemble_L3* dengan nilai *score_val* sebesar 0.939724 dan nilai *accuracy* yang dihasilkan pada model sebesar 0.94.

6.2 Saran

Berikut merupakan saran untuk penelitian selanjutnya berdasarkan penelitian yang telah dilaksankan:

1. Penelitian selanjutnya dapat melakukan eksplorasi *hyperparameter* lain secara optimal yang bertujuan untuk meningkatkan nilai *accuracy* dalam penelitian selanjutnya. Selain itu juga dapat menggunakan indikator yang lain dalam melakukan perbandingan mengani klasifikasi seperti *f1, precision,* maupun *recall*. Kemudian penelitian selanjutnya juga dapat melakukan prediksi dengan model regresi.

DAFTAR PUSTAKA

Abdurraman, G. (2023). Klasifikasi Kanker Payudara Menggunakan Algoritma SVM dengan Kernel RBF, Linier, dan Sigmoid. *Jurnal Sistem Informasi Ibrahimy*, 74~80.

Adi Nugroho, A. B. (2020). Perbandingan Performansi Algoritma Pengklasifikasian Terpandu Untuk Kasus Penyakit Kardiovaskular. *Jurnal Rekayasa Sistem dan Teknologi Informasi*, 998-1006.

Agus Ambarwari, Q. J. (2020). Analisis Pengaruh Data Scaling Terhadap Performa Algoritme Machine Learning untuk Identifikasi Tanaman. *Jurnal Rekayasa Sistem dan Teknologi Informasi*, 117-122.

Ahmadien Hafizh Yusufi, A. K. (2022). Prediksi Resiko Kematian Pada Penderita Penyakit Kadiovaskular Menggunakan Metode Ensemble Learning. *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya*, 531-541.

Alvaro Talavera, A. L. (2020). Machine Learning: A Contribution to Operational Research. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 1-7.

Amazon Sage Maker. (2023). Retrieved from AutoGluon-Tabular: https://docs.aws.amazon.com/sagemaker/latest/dg/autogluon-tabular.html

Amin Nur Rais, A. S. (2019). Integrasi SMOTE dan Ensemble AdaBoost Untuk Mengatasi Imbalance Class Pada Data Bank Direct Marketing. *Jurnal Informatika*, 278-285.

An Dinh, S. M. (2019). A Data-Driven Approach to Predicting. *BMC Medical Informatics nad Decision Making*, 1-15.

Annisa, R. (2019). Analisis Komprasi Algoritma Klasifikasi Data Mining untuk Prediksi Penderita Penyakit Jantung. *Jurnal Teknik Informatika Kaputama*, 22-28.

Apriyanto Alhamad, A. I. (2019). Prediksi Penyakit Jantung Menggunakan Metode-Metode Machine Learning Berbasis Ensemble – Weighted Vote. *Jurnal Edukasi dan Penelitian Informatika*, 352 - 360.

Ardana, A. (2023). Performance Analysis of XGBoost Algorithm to Determine the Most Optimal Parameters and Features in Predicting Stock Price Movement. *Telematika: Jurnal Informatika dan Teknologi Informasi*, 91-102.

Association, A. H. (2018). Heart Disease and Stroke Statistics - 2018 Update. *Aha Statistical Update*, e67-e492.

Association, A. H. (2022, Desember 6). *Understand Your Risks to Prevent a Heart Attack*. Retrieved from https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack

AutoGluon.Ai. (2023). Retrieved from TabularPredictor.leaderboard: https://auto.gluon.ai/stable/api/autogluon.tabular.TabularPredictor.leaderboard.html

AutoGluon.Ai. (2023). Retrieved from AutoGluon Tabular - Essential Functionality: https://auto.gluon.ai/stable/tutorials/tabular/tabular-essentials.html

AutoML.org. (2023). Retrieved from AutoML: https://www.automl.org/automl/

Ayu Novita Sari, S. A. (2022). Klasifikasi Penyakit Jantung Menggunakan Metode Naïve Bayes. *Analisa, Metode, Rekayasa, Informatika*, 22-26.

Bahrynovska, T. (2022, Maret 17). *Forbytes*. Retrieved from How Credit Scoring Software Solutions Help Assess Creditworthiness: https://forbytes.com/blog/credit-scoring-software/

Bhutani, K. (2023, September 29). *Geeks for geeks*. Retrieved from Pandas DataFrame describe() Method: https://www.geeksforgeeks.org/python-pandas-dataframe-describe-method/?ref=lbp

Budiharto, W. (2016). *Machine Learning dan Computational Intelligence.* Yogyakarta: Penerbit Andi.

CDC (Centers for Disease Control and Prevention). (2023, Mei 15). Retrieved from About Heart Disease: https://www.cdc.gov/heartdisease/about.htm

Cleveland Clinic. (2022). Retrieved from Cardiovascular Disease: https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease

Clinic, C. (2022, September 5). Retrieved from https://my.clevelandclinic.org/health/articles/9464-body-mass-index-bmi

Clinic, C. (2023, November 11). *Arthritis*. Retrieved from https://my.clevelandclinic.org/health/diseases/12061-arthritis

Costa, T., Silva, F., & Ferreira, L. P. (2017). Improve the extrusion process in tire production using Six Sigma methodology. *Procedia Manufacturing*, 1104-1111.

Detrinal Putra, A. W. (2020). Prediksi Keputusan Minat Penjurusan Siswa SMA Yadika 5 Menggunakan Algoritma Naïve Bayes. *Prosiding Seminar Nasional Riset Dan Information Science* , 84-92.

Devilia Rahmawati, E. P. (2020). Aplikasi Simpangan Baku Menggunakan Microsoft Excel. *Jurnal Matematika*, 47-53.

Dewi Cahyantia, A. R. (2020). Analisis performa metode Knnpada Dataset pasien pengidap Kanker Payudara. *Indonesian Journal of Data and Science*, 39 - 43.

Dewi, L. P. (2021, September 29). *Jenis, Gejala, dan Penyebab Penyakit Jantung*. Retrieved from Pemerintah Kota Surabaya RSUD dr. Mohammad Soewandhie: https://rs-soewandhi.surabaya.go.id/jenis-gejala-dan-penyebab-penyakit-jantung/#:~:text=Faktor%20Penyebab%20Resiko%20Penyakit%20Jantung&text=Faktor%20resiko%20yang%20tidak%20dapat%20dirubah%20antara%20lain%20usia%2C%20jenis,fisik%2C%20dan%20konsumsi%20alk

Dito Putro Utomo, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *Jurnal Media Informatika Budidarma,* 437 - 444.

Documentation, A. 1. (2023). *TabularPredictor.evaluate*. Retrieved from https://auto.gluon.ai/stable/api/autogluon.tabular.TabularPredictor.evaluate.html

Documentation, A. 1. (2023). *TabularPredictor.fit_weighted_ensemble*. Retrieved from https://auto.gluon.ai/stable/api/autogluon.tabular.TabularPredictor.fit_weighted_ensemble.html

Dwi Nugraheny, A. S. (2022). Analysis of the Validity of Determination of Graduation Predicate Based on Student Individual Data at Institut Teknologi Dirgantara Adisutjipto (ITDA). *Jurnal Multidisiplin Madani*, 1067 - 1082.

Eka Wahyudi, S. H. (2017). Case-Based Reasoning untuk Diagnosis Penyakit Jantung. *Icjs*, 1-10.

Fajar Sodik Pamungkas, B. D. (2020). Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python. *Prosiding Seminar Nasional Matematika*, 689-694.

Fangatulo Dodo, P. H. (2019). Penggunaan Machine Learning di Bidang Kesehatan. *Jurnal Penelitian Teknik Informatika*, 391-399.

Fathurohman, A. (2021). Machine Learning untuk Pendidikan : Mengapa dan Bagaimana. *Jurnal Informatika dan Teknologi Komputer*, 57-62.

Fayeza Sifat Fatima, D. A. (2022). Heart Disease Prediction Using Supervised Classifiers. *Proceedings of the International COnference on Innovative Computing & Communication*, 1-7.

Fienda Altamevia, H. O. (2023). Analisis Pola Penjualan Obat di Apotek Srikandi Menggunakan Algoritma Supervised Learning. *Jurnal Penerapan Sistem Informasi*, 170-176.

Fitri Handayani, K. S. (2021). Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network dalam Prediksi Penyakit Jantung. *Jurnal Edukasi dan Penelitian Informatika*, 329-334.

Hartina Hiromi Satyanegara, K. R. (2022). Implementation of CNN-MLP and CNN-LSTM for MitM Attack Detection System. *Rekayasa Sistem dan Teknologi Informasi*, 387-396.

Hasil Utama Riskesdas 2018. (2018). In K. K. Indonesia. Kementerian Kesehatan Republik Indonesia.

Hastuti, K. (2012). Analisis Komparasi Algoritma Klsaifikasi Data Mining untuk Prediksi Mahasiswa Non Aktif. *Seminar Nasional Teknologi Informasi & Komunikasi Terapan*, 241-249.

Heri Santoso, R. A. (2023). Deteksi Komentar Cyberbullying pada Media Sosial Instagram Menggunakan Algoritma Random Forest. *Jurnal Manajemen Informatika* , 62 - 72.

Hoiriyah, F. E. (2022). Prediksi Laju Pertumbuhan Jumlah Penduduk Provinsi Kalimantan Selatan Menggunakan Metode K-Nearest Neighbor Regression. *Technologia*, 351-355.

Ian H. Witten, E. F. (2006). *Data Mining : Practical Machine Learning Tools and Techniques-2nd Edition.* San Francisco: Morgan Kaufman Publishers.

Indonesia, K. K. (2020). Profil Kesehatan Indonesia Tahun 2019. Jakarta: Kemeterian Kesehatan Republik Indonesia.

Indriani, A. (2014). Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier. *Seminar Nasional Aplikasi Teknologi Informasi*, G-5 - G-10.

Irawan, M. I. (2004). Exploratory Data Analysis dengan JST - Kohonen SOM : Struktur Tingkat Kesejahteraan Daerah Tk II se Jawa Timur. *Seminar Nasional Aplikasi Teknologi Informasi*, J-27 - J-34.

Ishak, A. R. (2022). Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Correlation Matrix With Heatmap. *Jambura Journal of Electrical and Electronics Engineering*, 169 - 174.

Ivana Alhabib, A. F. (2022). Komparasi Metode Deep Learning, Naïve Bayes dan Random Forest untuk Prediksi Penyakit Jantung. *Informatics for Educators and Professionals*, 176-185.

Jung Hyun An, Z. W. (2023). A CNN-Based Automatic Vulnerability. *Journal on Wireless Communications and Networking*, 1-13.

Kanade, V. (2022, Agustus 30). *Spice Works*. Retrieved from What Is Machine Learning? Definition, Types, Applications, and Trends for 2022 : https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/amp/

Kanghyeon Seo, B. C. (2021). Forecasting the Walking Assistance Rehabilitation Level of Stroke Patients Using Artificial Intelligence. *MDPI*, 1-17.

Kolondam, Y. K. (2021, Desember 23). *Gamelab Indonesia*. Retrieved from Mengenal Machine Learning: Pengertian, Jenis dan Cara Kerja: https://www.gamelab.id/news/1292-mengenal-machine-learning-pengertian-jenis-dan-cara-kerja

Krishnan, A. (2020, Januari 9). *Amazon Sciene*. Retrieved from Amazon’s AutoGluon Helps Developers Deploy Deep Learning Models with Just A Few lines of Code: https://www.amazon.science/amazons-autogluon-helps-developers-get-up-and-

running-with-state-of-the-art-deep-learning-models-with-just-a-few-lines-of-code

Kualitatif, S. D. (2016). Amirotun Sholikhah. *Komunika*, 342-362.

Laila Qadrini, A. S. (2021). Decision Tree dan Adaboost pada Klasifikasi Penerima Program Bantuan Sosial. *Jurnal Inovasi Penelitian*, 1959 - 1966.

Leonita Angelina, Y. P. (2016). Pengambilan Keputusan Pada Trafik Management Dengan Menggunakan Reinforcement Learning. *e-Proceeding of Engineering : Vol.3*, 4964-4971.

Lestari, D. (2022). Metode Naive Bayes dalam Machine Learning untuk Memprediksi Penyakit Jantung dalam Tubuh. *Jurnal Teknologi Komputer dan Sistem Informasi*, 23-28.

Levande, V. (2021, Oktober 6). *Analytics In Diamag*. Retrieved from A Guide to Using AutoGluon for Automating Machine Learning Tasks: https://analyticsindiamag.com/a-guide-to-using-autogluon-for-automating-machine-learning-tasks/

Lucas B.V. de Amorima, G. D. (2022). The Choice of Scaling Technique Matters for Classification Performance. *Arxiv*, 1-37.

Luthfiana Ratnawati, D. R. (2019). Penerapan Random Forest untuk Mengukur Tingkat Keparahan Penyakit pada Daun Apel. *Jurnal Sains dan Seni Its*, A71 - A77.

Luthfiyah Amatullah, Y. W. (2022). Penerapan Klasifikasi Random Forest Terhadap Data Gangguan Spektrum Autisme (ASD) Pada Anak – Anak Menggunakan Seleksi Fitur Principal Component Analysis. *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya*, 659-667.

M. Azwan, R. K. (2021). Penerapan Algoritma K-Means Clustering dan Correlation Matrix Untuk Menganalisis Risiko Penyebaran Demam Berdarah di Kota Pekanbaru. *Jurnal Informatika Merdeka Pasuruan*, 1-6.

M. Swathy, K. S. (2022). A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. *Ict Express 8*, 109-116.

Manzilur Rahman Romadhon, F. K. (2021). A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia. *East Indonesia Conference on Computer and Information Technology*, 41-44.

Mardhiyatirrahmah, L. (2023). Pembelajaran Statistika Terkait Ukuran Pemusatan Data (Mean, Modus, dan Median) melalui Integrasi Terhadap Al-Qur'an. *Jurnal Ilmiah Ilmu Kependidikan dan Kedakwahan*, 41-50.

Mawadatul Maulidah, W. G. (2020). Algoritma Klasifikasi Decision TreeUntuk Rekomendasi Buku Berdasarkan Kategori Buku. *Jurnal Ilmih Ekonomi dan Bisnis*, 89-96.

Maximillian Christianto, J. A. (2020). Aplikasi Analisa Sentimen Pada Komentar Berbahasa Indonesia Dalam Objek Video di Website YouTube Menggunakan Metode Naïve Bayes Classifier. *Jurnal Infra*, 255-259.

Md Mamun Ali, B. K. (2021). Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis and Comparison. *Computers in Biology and Medicine*, 1-12.

Mitchell, T. M. (1997). *Machine Learning.* New York: McGraw-Hill Science.

Moch Haris, N. F. (2023). Mengoptimalkan Belanja Operasional di Badan Perencanaan Eselon I terhadap Total Belanja Guna Mewujudkan Organisasi yang Kuat dan Profesional. *Journal of Industrial Engineering & Management Research*, 71-92.

Mochammad Anshori, N. R. (2022). Prediksi Pasien dengan Penyakit Kardiovaskular Menggunakan Random Forest. *Jurnal Tektrika*, 58-64.

Mohamad Adhisyanda Aditya, R. D. (2020). Penggabungan Teknologi Untuk Analisa Data Berbasis Data Science. *Seminar Nasional Teknologi Komputer & Sains*, 51-56.

Muhammad Raffi, A. S. (2023). Analisis Sentimen Ulasan Aplikasi Binar Pada Google Play Store Menggunakan Algoritma Naive Bayes. *Journal of Information Technology and Computer Sciene*, 238-462.

Murdifin Haming, R. S. (2019). *Operation Research : Teknik Pengambilan Keputusan Optimal.* Jakarta: Bumi Aksara.

NHS UK. (2022, April 22). Retrieved from Cardiovascular disease: https://www.nhs.uk/conditions/cardiovascular-disease/

Nick Erickson, J. M. (2020). AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *Arxiv*, 1-28.

Novia Hasdyna, R. K. (2020). Analisis Matthew Correlation Coefficientpada K-Nearest Neighbor dalam Klasifikasi Ikan Hias. *Informatics Journal*, 57-64.

Nugraha, W. (2021). Prediksi Penyakit Jantung Cardiovascular Menggunakan Model Algoritma Klasifikasi. *Jurnal Sigma*, 78-84.

Nur Baiti Nasution, D. H. (2023). Prediksi Lama Studi dan Predikat Kelulusan Mahasiswa Menggunakan Algoritma Supervised Learning. *Jurnal Teknologi Terapan*, 386-395.

Obaid Alotaibi, E. P. (2023). Cleaning Big Data Streams: A Systematic Literature Review. *MDPI*, 1-24.

Oleksandr Shchur, C. T. (2023). AutoGluon–TimeSeries: AutoML for Probabilistic Time Series Forecasting. *Arxiv*, 1-21.

Oluwaseyi Ogunfowora, H. N. (2023). Reinforcement and Deep Reinforcement Learning-Based Solutions for Machine Maintenance Planning, Scheduling Policies, and Optimization. *Journal of Manufacturing Systems*, 244-263.

Prasanna, S. (2020, Maret 31). *Machine Learning with AutoGluon, an open source AutoML library*. Retrieved from AWS Open Source Blog: https://aws.amazon.com/id/blogs/opensource/machine-learning-with-autogluon-an-open-source-automl-library/

Prevention, C. f. (2023, Maret 21). *Know Your Risk for Heart Disease*. Retrieved from https://www.cdc.gov/heartdisease/risk_factors.htm

Profil Kesehatan Indonesia 2021. (2022). Jakarta: Kemeterian Kesehatan Republik Indonesia.

Puput Santoso, H. A. (2021). Algoritma Supervised Learning dan Unsupervised Learing dalam Pengolahan Data. *G-Tech : Jurnal Teknologi Terapan*, 315-318.

Putra, J. W. (2020). *Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.4*. Tokyo.

Raja Krishnamoorthi, S. J. (2022). A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *Hindawi*, 1-10.

Rajkumar Gangappa Nadakinamani, A. S. (2022). Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques. *Hindawi*, 1-12.

Ranjan, S. (2021, Agustus 25). *Geeks for geeks*. Retrieved from Python | Pandas dataframe.info(): https://www.geeksforgeeks.org/python-pandas-dataframe-info/

Reni Amelia, A. M. (2022). Regresi Logistik Biner dengan Proses Resamplingdalam Menduga Faktor Determinan Merokok Remaja. *Seminar Nasional Official Statistics*, 703-712.

Rerung, R. R. (2018). Penerapan Data Mining dengan Memanfaatkan Metode Association Rule untuk Promosi Produk. *Jurnal Teknologi Rekayasa*, 89-98.

Rifqi Kurniawan, I. C. (2015). Ensemble Machine Learning untuk Rekomendasi Penilaian Kinerja Guru Berbasis Weighted Product (Studi Kasus Sekolah Dasar di Kecamatan Rogojampi. *DORO: Repository Jurnal Mahasiswa PTIIK Universitas Brawijaya*, 1-12.

Saifullah, M. Z. (2017). Analisa Terhadap Perbandingan Algoritma Decision Tree dengan Algoritma Random Tree untuk Pre-Processing Data. *Jurnal Sains Komputer dan Informatika*, 180-185.

Sanjiv R. Da, M. D. (2022). FinLex: An Effective Use of Word Embeddings for Financial Lexicon Generation. *The Journal of Finance and Data Science*, 1-11.

Sarbaini, E. P. (2021). Pengelompokan Diabetic Macular Edema Berbasis Citra Retina Mata Menggunakan Fuzzy Learning Vector Quantization (FLVQ). *Jurnal Sains, Teknologi dan Industri*, 75 - 80.

Sri Sumarlinda, W. L. (2022). Aplikasi K-Nearest Neighbor (KNN) untuk Klasifikasi Penyakit Kardiovaskuler. *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, 259-261.

Taghsya Izmi Andini, W. W. (2016). Prediksi Potensi Pemasaran Produk Baru dengan Metode Naïve Bayes Classifier dan Regresi Linear. *Seminar Nasional Aplikasi Teknologi Informasi* , A-27 - A-32.

Tan, H. T. (2012). Metode DMAIC Sebagai Solusi Pengendalian Kualitas Produksi Sepatu Tambang:Studi Kasus PT Mangul Jaya-Bekasi. *ComTech, 3*, 509-523.

UK, N. (2022, November 28). *What is the body mass index (BMI)?* Retrieved from https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/

Urminder Singh, M. H. (2020). MetaOmGraph: A Workbench for Interactive Exploratory Data Analysis of Large Expression Datasets. *Nucleic Acids Research*, 1-19.

Vercellis, C. (2009). *Business Intelligence : Data Mining and Optimization for Decision Making.* Milan: Wiley.

Vibhutijain. (2021, November 22). *Training vs Testing vs Validation Sets*. Retrieved from Geeks for geeks: https://www.geeksforgeeks.org/training-vs-testing-vs-validation-sets/?ref=gcse

Vitho, I., Ginting, E., & Anizar. (2013). Aplikasi Six Sigma Untuk Menganalisis Faktor-faktor Penyebab Kecacatan Produk Crumb Rubber Sir 20 Pada Pt. XYZ. *e-Jurnal Teknik Industri FT USU Vol 3, No. 4*, 23-28.

Weicheng Sun, P. Z. (2021). Prediction of Cardiovascular Diseases based on Machine Learning. *Asp Transactions on Internet of Things*, 30-35.

Wen Zhu, N. Z. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, 1-9.

Wenwen Qi, C. X. (2021). AutoGluon: A Revolutionary Framework for Landslide Hazard Analysis. *Natural Hazards Research*, 103-108.

Wiji Lestari, S. S. (2023). Studi Komparatif Model Klasifikasi Kerentanan Penyakit Jantung Menggunakan Algoritma Machine Learning. *Sains dan Teknologi Informasi*, 107 - 115.

Wilem Musu, A. I. (2021). Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5. *Prosiding Seminar Ilmiah Sistem Informasi dan Teknologi Informasi*, 186-195.

Wisnubroto, P., & Rukmana, A. (2015). Pengendalian Kualitas Produk dengan Pendekatan Six Sigma dan Analisis Kaizen serta New Seven Tools Sebagai Usaha Pengurangan Kecacatan Produk. *Jurnal Teknologi*, 65-74.

World Health Organization. (2019). Retrieved from Health Data Overview for the Republic of Indonesia: https://data.who.int/countries/360

World Health Organization. (2021, Juni 11). Retrieved from Cardiovascular diseases (CVDs): https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

World Heart Federation. (2023). Retrieved from What Is Cardiovascular Disease?: https://world-heart-federation.org/what-is-cvd/

Yosiko Aditya Pratama, F. B. (2023). Analisis Optimasi Algoritma Decision Tree, Logistic Regression dan SVM Menggunakan Soft Voting. *Jurnal Media Informatika Budidarma*, 1908-1919.

Yuli Sun Hariyani, S. H. (2020). Deteksi Penyakit Covid-19 Berdasarkan Citra X-Ray Menggunakan Deep Residual Network . *Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika* , 443-453.

Yuliana, Nasution, Y. N., & Wasono. (2017). Penggunaan Metode Kaizen Pada Tahap Improve Dalam Six Sigma (Studi Kasus: Perusahaan Air Minum Dalam Kemasan(AMDK) Merk RAMA Produksi PT Ranam Mahakam Indonesia). *Jurnal Eksponensial*.

Zhongming Wu, Y. S. (2023). Data-Driven Distributionally Robust Support Vector Machine Method for Multiple Criteria Sorting Problem with Uncertainty. *Journal Pre-proof*, 1-28.

Zoran Stojanoski, M. K. (2023). Comparative Analysis of Machine Learning Models for Diabetes Prediction . *Proceedings of the 11th International Conference on Applied Innovations in IT*, 75-80.

LAMPIRAN

A- Lampiran Pemodelan *AutoGluon*

```
#Pre-processing Data
##Installasi dan Import Modul
#Melakukan import library
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from imblearn.over_sampling import SMOTE
#Melakukaan installasi library autogluon
!pip install autogluon
#Melakukan installasi modul tabular
!pip install autogluon.tabular
# Mengunistall scikit-learn versi terkini
!pip uninstall scikit-learn -y
# Menginstall scikit-learn versi yang diinginkan
!pip install scikit-learn==1.2.2
#Melakukan import library autogluon
import autogluon as ag
from autogluon.tabular import TabularPredictor as task
#Melakukan import dataset
data = pd.read_csv('/content/drive/MyDrive/TA/CVD_cleaned.csv')
#Menyalin data
data1 = data.copy()
#Menampilkan 5 data teratas
```

```
data.head()
#Menapilkan 5 data terbawah
data.tail()
##Informasi Data
#Menampilkan informasi jumlah dan jenis data
data.info()
##EDA (Exploratory Data Analysis)
#Mendefinisikan variabel kategori berdasarkan tipe kolom data
variabel_kategori = [var for var in data.select_dtypes(include=['object', 'float64']).columns]
#Membuat visualisasi dengan subplot
jumlah_kolom = len(variabel_kategori)
jumlah_baris = (jumlah_kolom + 3) // 4
#nrows= jumlah_baris
#ncols = jumlah_kolom
fig, axs = plt.subplots(nrows=jumlah_baris, ncols=4, figsize=(20, 7*jumlah_baris))
axs = axs.flatten()
#Menampilkan 5 nilai teratas pada variabel kategori dengan histogram
for i, var in enumerate(variabel_kategori):
  nilai_teratas = data[var].value_counts().nlargest(5).index
  data_filter = data[data[var].isin(nilai_teratas)]
  sns.countplot(x=var, data=data_filter, ax=axs[i])
  axs[i].set_title(var)
  axs[i].tick_params(axis='x', rotation=90)
#Menghapus subplot kosong yang tidaK diperlukan
if jumlah_kolom < len(axs):
    for i in range(jumlah_kolom, len(axs)):
        fig.delaxes(axs[i])
#Menampilkan hasi visualisasi
fig.tight_layout()
plt.show()
#Mendefinisikan jumlah data
jumlah_data = pd.DataFrame(data)
```

```
#Menampilkan nilai data secara keseluruhan
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
#Menampilkan nilai masing-masing variabel
for column in jumlah_data.columns:
    value_counts = jumlah_data[column].value_counts()
    print(f"Jumlah nilai pada variabel '{column}':")
    print(value_counts)
    print()
##Menampilkan hubungan Variabel Heart Disease dengan Variabel yang lainnya
#Mengelompokan semua variabel selain variabel Heart Disease menjadi satu kelompok
semua_variabel = ['General_Health', 'Checkup', 'Exercise', 'Skin_Cancer', 'Other_Cancer', 'Depression', 'Diabetes', 'Arthritis', 'Sex', 'Age_Category', 'Height_(cm)', 'Weight_(kg)', 'BMI', 'Smoking_History', 'Alcohol_Consumption', 'Fruit_Consumption', 'Green_Vegetables_Consumption', 'FriedPotato_Consumption']
#Menampilkan hubungan antara variabel Heart Disease dengan variabel lainnya
for variabel in semua_variabel:
  plt.figure(figsize=(12,5))
  sns.countplot(data=data, x= variabel, hue='Heart_Disease')
  plt.title('Tampilan hubungan antara ' +  variabel  + ' dengan Heart Disease')
  plt.xticks(rotation=0)
  plt.show()
#Mengenlompokan variabel dengan nilai non numerical
variabel_non_numerical = ['General_Health', 'Checkup', 'Exercise', 'Skin_Cancer', 'Other_Cancer', 'Depression', 'Diabetes', 'Arthritis', 'Sex', 'Smoking_History']
#Menampilkan jumlah variabel non numerical dengan variabel heart disease
for non_numerical in variabel_non_numerical:
                    jumlah_variabel = data.groupby([non_numerical, 'Heart_Disease']).size().reset_index(name='Count')
    print(f"\Jumlah Variabel {non_numerical} yang berhubungan dengan Variabel Heart Disease:")
    print(jumlah_variabel)
#Menampilkan jumlah variabel Age Category  dengan variabel heart disease
```

```
jumlah_variabel_age_category = data.groupby(['Age_Category', 'Heart_Disease']).size().reset_index(name='Count')
print("Jumlah Variabel Age_Category yang berhubungan dengan Variabel Heart Disease:")
print(jumlah_variabel_age_category)
#Nilai modus pada hubungan variabel Age Category dengan variabel heart disease
modus_age_category = data['Age_Category'].mode()
print('Nilai modus variabel Age Category :', modus_age_category[0])
#Jumlah nilai modus pada variabel Age Category
jumlah_modus_age_category = data[data['Age_Category'] == modus_age_category[0]].shape[0]
print("Jumlah nilai Modus Variabel Age Category:", jumlah_modus_age_category)
#Menampilkan jumlah variabel Height (cm) dengan variabel heart disease
jumlah_variabel_height_cm = data.groupby(['Height_(cm)', 'Heart_Disease']).size().reset_index(name='Count')
print("Jumlah Variabel Height_(cm) yang berhubungan dengan Variabel Heart Disease:")
print(jumlah_variabel_height_cm)
#Nilai modus pada hubungan variabel Height (cm) dengan variabel heart disease
modus_height_cm = data['Height_(cm)'].mode()
print('Nilai modus variabel Height (cm) :', modus_height_cm[0])
#Jumlah nilai modus pada variabel Height (cm)
jumlah_modus_height_cm = data[data['Height_(cm)'] == modus_height_cm[0]].shape[0]
print("Jumlah nilai Modus Variabel Height (cm):", jumlah_modus_height_cm)
#Menampilkan jumlah variabel Weight (kg) dengan variabel heart disease
jumlah_variabel_weight_kg = data.groupby(['Weight_(kg)', 'Heart_Disease']).size().reset_index(name='Count')
print("Jumlah Variabel Weight_(kg) yang berhubungan dengan Variabel Heart Disease:")
print(jumlah_variabel_weight_kg)
#Nilai modus pada hubungan variabel Weight (kg) dengan variabel heart disease
modus_weight_kg = data['Weight_(kg)'].mode()
print('Nilai modus variabel Weight (kg) :', modus_weight_kg[0])
#Jumlah nilai modus pada variabel Weight (kg)
```

```
jumlah_modus_weight_kg = data[data['Weight_(kg)'] == modus_weight_kg[0]].shape[0]
print("Jumlah nilai Modus Variabel Weight (kg):", jumlah_modus_weight_kg)
#Menampilkan jumlah variabel BMI dengan variabel heart disease
jumlah_variabel_bmi = data.groupby(['BMI', 'Heart_Disease']).size().reset_index(name='Count')
print("Jumlah Variabel BMI yang berhubungan dengan Variabel Heart Disease:")
print(jumlah_variabel_bmi.head(681))
#Menampilkan jumlah variabel BMI dengan variabel heart disease
jumlah_variabel_bmi = data.groupby(['BMI', 'Heart_Disease']).size().reset_index(name='Count')
pd.set_option('display.min_rows', None)
pd.set_option('display.max_columns', None)
print("Jumlah Variabel BMI yang berhubungan dengan Variabel Heart Disease:")
print(jumlah_variabel_bmi)
#Nilai modus pada hubungan variabel BMI dengan variabel heart disease
modus_bmi = data['BMI'].mode()
print('Nilai modus variabel BMI :', modus_bmi[0])
#Jumlah nilai modus pada variabel BMI
jumlah_modus_bmi = data[data['BMI'] == modus_bmi[0]].shape[0]
print("Jumlah nilai Modus Variabel BMI:", jumlah_modus_bmi)
#Menampilkan jumlah variabel Alcohol Consumption dengan variabel heart disease
jumlah_variabel_alcohol_consumption = data.groupby(['Alcohol_Consumption', 'Heart_Disease']).size().reset_index(name='Count')
print("Jumlah Variabel Alcohol_Consumption yang berhubungan dengan Variabel Heart Disease:")
print(jumlah_variabel_alcohol_consumption)
#Nilai modus pada hubungan variabel Alcohol Consumption dengan variabel heart disease
modus_alcohol_consumption = data['Alcohol_Consumption'].mode()
print('Nilai modus variabel Alcohol_Consumption :', modus_alcohol_consumption[0])
#Jumlah nilai modus pada variabel Alcohol Consumption
```

```
jumlah_modus_alcohol_consumption = data[data['Alcohol_Consumption'] == modus_alcohol_consumption[0]].shape[0]
print("Jumlah nilai Modus Variabel Alcohol_Consumption:", jumlah_modus_alcohol_consumption)
#Menampilkan jumlah variabel Fruit Consumption dengan variabel heart disease
jumlah_variabel_fruit_consumption = data.groupby(['Fruit_Consumption', 'Heart_Disease']).size().reset_index(name='Count')
print("Jumlah Variabel Fruit_Consumption yang berhubungan dengan Variabel Heart Disease:")
print(jumlah_variabel_fruit_consumption)
#Nilai modus pada hubungan variabel Fruit Consumption dengan variabel heart disease
modus_fruit_consumption = data['Fruit_Consumption'].mode()
print('Nilai modus variabel Fruit Consumption :', modus_fruit_consumption[0])
#Jumlah nilai modus pada variabel Fruit Consumption
jumlah_modus_fruit_consumption = data[data['Fruit_Consumption'] == modus_fruit_consumption[0]].shape[0]
print("Jumlah nilai Modus Variabel Fruit Consumption:", jumlah_modus_fruit_consumption)
#Menampilkan jumlah variabel Green Vegetables Consumption dengan variabel heart disease
jumlah_variabel_green_vegetable_consumption = data.groupby(['Green_Vegetables_Consumption', 'Heart_Disease']).size().reset_index(name='Count')
print("Jumlah Variabel Green_Vegetables_Consumption yang berhubungan dengan Variabel Heart Disease:")
print(jumlah_variabel_green_vegetable_consumption)
#Nilai modus pada hubungan variabel Green Vegetables Consumption dengan variabel heart disease
modus_green_vegetables_consumption = data['Green_Vegetables_Consumption'].mode()
print('Nilai modus variabel Green Vegetables Consumption :', modus_green_vegetables_consumption[0])
#Jumlah nilai modus pada variabel Green Vegetables Consumption
```

```
jumlah_modus_green_vegetables_consumption =
data[data['Green_Vegetables_Consumption'] ==
modus_green_vegetables_consumption[0]].shape[0]
print("Jumlah nilai Modus Variabel Green Vegetables Consumption :",
jumlah_modus_green_vegetables_consumption)
#Menampilkan jumlah variabel Fried Potato Consumption dengan variabel heart disease
jumlah_variabel_fried_potato_consumption =
data.groupby(['FriedPotato_Consumption',
'Heart_Disease']).size().reset_index(name='Count')
print("Jumlah Variabel FriedPotato_Consumption yang berhubungan dengan Variabel
Heart Disease:")
print(jumlah_variabel_fried_potato_consumption)
#Nilai modus pada hubungan variabel Fried Potato Consumption dengan variabel heart
disease
modus_fried_potato_consumption = data['FriedPotato_Consumption'].mode()
print('Nilai modus variabel Fried Potato Consumption :',
modus_fried_potato_consumption[0])
#Jumlah nilai modus pada variabel Green Vegetables Consumption
jumlah_modus_fried_potato_consumption = data[data['FriedPotato_Consumption'] ==
modus_fried_potato_consumption[0]].shape[0]
print("Jumlah nilai Modus Variabel Fried Potato Consumption :",
jumlah_modus_fried_potato_consumption)
##Memeriksa Kondisi Data
#Memeriksa data isnull
data.isnull()
#Memeriksa data duplikat
data_duplikat = data[data.duplicated()]
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
print(data_duplikat)
##Encoding Data
#Melakukan konversi nilai Label Encoding
from sklearn.preprocessing import LabelEncoder
```

```
kategori_label_encoding = data[['General_Health','Checkup', 'Diabetes', 'Age_Category','Exercise', 'Heart_Disease', 'Skin_Cancer', 'Other_Cancer', 'Depression', 'Arthritis', 'Sex', 'Smoking_History']].columns
#Memulai konversi nilai label encoding
LE = LabelEncoder()
for i in kategori_label_encoding:
  data[i] = LE.fit_transform(data[i])
data.head()
data.tail()
##Data Cleaning
###Handling Missing Values
#Memeriksa jumlah data yang hilang
data.isnull().sum()
###Handling Dupplicate
# Menjumlahkan data yang terduplikat
duplicate_count = data.duplicated().sum()
# Mencetak jumlah data yang terduplikat
print("Jumlah data yang terduplikat sebanyak:", duplicate_count)
# Menghapus data yang terduplikat
data = data.drop_duplicates()
###Feature Scaling
#Mendefinisikan Variabel
X = data.iloc[:, 1:-1].values
Y = data.iloc[:, -1].values
Y = Y.reshape(-1, 1)
#Mendefinisikan Standard Scaler
from sklearn.preprocessing import StandardScaler
sc_x = StandardScaler()
sc_y = StandardScaler()
x = sc_x.fit_transform(X)
y = sc_y.fit_transform(Y)
#Menampilkan variabel X
print(X)
```

```
#Menampilkan variabel Y
print(Y)
##Matriks Korelasi
#Menampilkan visualisasi matriks dengan menggunakan heatmap
plt.figure(figsize=(15, 15))
sns.heatmap(data.corr(), annot=True, cmap="YlGnBu")
plt.title('Matriks Korelasi antar variabel')
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
plt.tight_layout()
plt.show()
##Deskripsi Data
#Menampilkan deskripsi statistik data dengan tranpose
data.describe().T
##Data Training dan Data Testing
#Dengan SMOTE
#Menetapkan kolom yang akan diprediksi
X = data
Y = data['Heart_Disease']
#Mendefinisikan smote
smote = SMOTE(random_state=0)
X_resample, Y_resample = smote.fit_resample(X, Y)
#Membagi antara data training dan data testing dengan perbandingan 7:3
X_train, X_test, Y_train, Y_test = train_test_split(X_resample, Y_resample,
test_size=0.3, random_state=0)
#Menampilkan hasil data training dan data testing
print('Jumlah data trainig sebanyak:', len(X_train))
print('Jumlah data testing sebanyak:', len(X_test))
#Tanpa SMOTE
#Menetapkan kolom yang akan diprediksi
X1 = data1
Y1 = data1['Heart_Disease']
```

```
#Membagi antara data training dan data testing dengan perbandingan 7:3
X1_train, X1_test, Y1_train, Y1_test = train_test_split(X1, Y1, test_size=0.3, random_state=0)
#Menampilkan hasil data training dan data testing
print('Jumlah data trainig sebanyak:', len(X1_train))
print('Jumlah data testing sebanyak:', len(X1_test))
#Membaca dataset data training dan data testing
test_data = pd.concat([X_test, Y_test], axis=1)
train_data = pd.concat([X_train, Y_train], axis=1)
#Pemodelan AutoGluon
#Installasi Tabular Predictor
predictor = task(label='Heart_Disease', problem_type = 'binary', eval_metric = 'accuracy', path = 'AutoGluonModels').fit(train_data = X_train, time_limit = 300, presets = 'best_quality')
#Menampilkan hasil pemeringkatan model autogluon
predictor.leaderboard()
#Mencari model terbaik
model_terbaik = predictor.get_model_best()
print(model_terbaik)
#Membuat visualisasi pada predictor
f, ax = plt.subplots(figsize = (8,4))
sns.barplot(x = 'model', y = 'score_val', data = predictor.leaderboard(), color = 'g')
ax.set(ylabel = 'score_val', xlabel = 'Model')
plt.xticks(rotation = 45);
predictor.evaluate(X_train)
#Evaluasi Model
#Melakukan perhitungan prediksi pada data testing
Y_pred = predictor.predict(X_test)
#Membuat confusion matrix pada data testing
conf_matrix = confusion_matrix(Y_test, Y_pred)
#Melakukan normalisasi pada confusion matrix dengan satuan persen
conf_matrix_perc = conf_matrix.astype('float')/conf_matrix.sum(axis=1)[:, np.newaxis]
*100
```

```
display_labels = ['Tidak menderita penyakit jantung', 'Menderita Penyakit Jantung']
label_values = [['{:.0f}\n({:.1f}%)'.format(val, perc) for val, perc in zip(row, row_perc)]
for row, row_perc in zip(conf_matrix, conf_matrix_perc)]
# Membuat heatmap dengan label-label
plt.figure(figsize=(8, 6))
heatmap = sns.heatmap(conf_matrix_perc, annot=label_values, fmt='',
cmap='nipy_spectral', cbar=True, xticklabels=display_labels,
yticklabels=display_labels)
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix')
# Menambahkan color scale legend
cbar = heatmap.collections[0].colorbar
cbar.set_label('Percentage')
plt.show()
```

B- Lampiran Hasil Model *AutoGluon*

```
#Melakukan import library
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from imblearn.over_sampling import SMOTE
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
#Melakukan import dataset
data = pd.read_csv('/content/drive/MyDrive/TA/CVD_cleaned.csv')
#Menampilkan 5 data teratas
data.head()
#Melakukan konversi nilai Label Encoding
kategori_label_encoding = data[['General_Health','Checkup', 'Diabetes',
'Age_Category','Exercise', 'Heart_Disease', 'Skin_Cancer', 'Other_Cancer', 'Depression',
'Arthritis', 'Sex', 'Smoking_History']].columns
#Memulai konversi nilai label encoding
LE = LabelEncoder()
for i in kategori_label_encoding:
  data[i] = LE.fit_transform(data[i])
data.head()
#Memeriksa jumlah data yang hilang
data.isnull().sum()
```

```
# Menjumlahkan data yang terduplikat
duplicate_count = data.duplicated().sum()
# Mencetak jumlah data yang terduplikat
print("Jumlah data yang terduplikat sebanyak:", duplicate_count)
# Menghapus data yang terduplikat
data = data.drop_duplicates()
#Mendefinisikan Variabel
X = data.iloc[:, 1:-1].values
Y = data.iloc[:, -1].values
Y = Y.reshape(-1, 1)
#Mendefinisikan Standard Scaler
from sklearn.preprocessing import StandardScaler
sc_x = StandardScaler()
sc_y = StandardScaler()
x = sc_x.fit_transform(X)
y = sc_y.fit_transform(Y)
#Menampilkan variabel X
print(X)
#Menampilkan variabel Y
print(Y)
#Dengan SMOTE
#Menetapkan kolom yang akan diprediksi
X = data.drop('Heart_Disease', axis=1)
Y = data['Heart_Disease'
#Mendefinisikan smote
smote = SMOTE(random_state=0)
X_resample, Y_resample = smote.fit_resample(X, Y)
#Membagi antara data training dan data testing dengan perbandingan 7:3
X_train, X_test, Y_train, Y_test = train_test_split(X_resample, Y_resample,
test_size=0.3, random_state=0)
#Menampilkan hasil data training dan data testing
print('Jumlah data trainig sebanyak:', len(X_train))
print('Jumlah data testing sebanyak:', len(X_test))
```

```
#Membuat model Random Forest
rf_model = RandomForestClassifier(n_estimators=100, random_state=0)
#Training model Random Forest
rf_model.fit(X_train, Y_train)
#Menghitung nilai akurasi model Random Forest
nilai_rf = rf_model.score(X_test, Y_test)
print("Nilai Akurasi  Random Forest: {:.2f}".format(nilai_rf))
#Membuat prediksi dengan model Random Forest
y_pred_rf = rf_model.predict(X_test)
report_rf = classification_report(Y_test, y_pred_rf)
print("Hasil Klasifikasi Random Forest:")
print(report_rf)
#Membuat model k-NN
knn_model = KNeighborsClassifier(n_neighbors=3)
#Training model k-NN
knn_model.fit(X_train, Y_train)
#Menghitung nilai akurasi model k-NN
nilai_knn = knn_model.score(X_test, Y_test)
print("Nilai Akurasi K-NN: {:.2f}".format(nilai_knn))
#Membuat prediksi dengan model k-NN
y_pred_knn = knn_model.predict(X_test)
report_knn = classification_report(Y_test, y_pred_knn)
print("Hasil Klasifikasi k-NN:")
print(report_knn)
#Mendefinisikan model Weighted Ensemble Learning dengan menggambungkan model
Random Forest dan K-NN
weighted_ensemble = {'random_forest': rf_model, 'k_nn': knn_model}
weights = {'random_forest': 0.7, 'k_nn': 0.3
#Menghitung prediksi Weighted Ensemble Learning
y_pred_weighted = None
for model_name, model in weighted_ensemble.items():
    if y_pred_weighted is None:
        y_pred_weighted = model.predict_proba(X_test) * weights[model_name]
```

```
    else:
        y_pred_weighted += model.predict_proba(X_test) * weights[model_name]
y_pred_weighted = y_pred_weighted.argmax(axis=1)
#Menghitung nilai akurasi Weighted Ensemble Learning
nilai_weighted = accuracy_score(Y_test, y_pred_weighted)
print("Nilai Akurasi Weighted Ensemble Learning: {:.2f}".format(nilai_weighted))
#Membuat prediksi dengan model Weighted Ensemble Learning
report_weighted = classification_report(Y_test, y_pred_weighted)
print("Hasil Klasifikasi Weighted Ensemble Learning:")
print(report_weighted)
# Melakukan perhitungan prediksi pada data testing Random Forest
y_pred_rf = rf_model.predict(X_test)
# Membuat confusion matrix pada data testing Random Forest
conf_matrix = confusion_matrix(Y_test, y_pred_rf)
# Melakukan normalisasi pada confusion matrix dengan satuan persen
conf_matrix_perc = conf_matrix.astype('float') / conf_matrix.sum(axis=1)[:, np.newaxis]
* 100
display_labels = ['Tidak menderita penyakit jantung', 'Menderita Penyakit Jantung']
label_values = [['{:.0f}\n({:.1f}%)'.format(val, perc) for val, perc in zip(row, row_perc)]
for row, row_perc in zip(conf_matrix, conf_matrix_perc)]
# Membuat heatmap dengan label-label
plt.figure(figsize=(8, 6))
heatmap = sns.heatmap(conf_matrix_perc, annot=label_values, fmt='',
cmap='nipy_spectral', cbar=True, xticklabels=display_labels,
yticklabels=display_labels)
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix  Random Forest')
plt.text(1, -0.5, 'Random Forest Model', fontsize=12, ha='center', va='center',
bbox=dict(facecolor='lightgray', edgecolor='gray', boxstyle='round,pad=0.5'))
# Menambahkan color scale legend
cbar = heatmap.collections[0].colorbar
cbar.set_label('Percentage')
```

```
plt.show()
# Melakukan perhitungan prediksi pada data testing k-NN
y_pred_knn = knn_model.predict(X_test)
# Membuat confusion matrix pada data testing k-NN
conf_matrix = confusion_matrix(Y_test, y_pred_knn)
# Melakukan normalisasi pada confusion matrix dengan satuan persen
conf_matrix_perc = conf_matrix.astype('float') / conf_matrix.sum(axis=1)[:, np.newaxis]
* 100
display_labels = ['Tidak menderita penyakit jantung', 'Menderita Penyakit Jantung']
label_values = [['{:.0f}\n({:.1f}%)'.format(val, perc) for val, perc in zip(row, row_perc)]
for row, row_perc in zip(conf_matrix, conf_matrix_perc)]
# Membuat heatmap dengan label-label
plt.figure(figsize=(8, 6))
heatmap = sns.heatmap(conf_matrix_perc, annot=label_values, fmt='',
cmap='nipy_spectral', cbar=True, xticklabels=display_labels,
yticklabels=display_labels)
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix  k-NN')
plt.text(1, -0.5, 'k-NN Model', fontsize=12, ha='center', va='center',
bbox=dict(facecolor='lightgray', edgecolor='gray', boxstyle='round,pad=0.5'))
# Menambahkan color scale legend
cbar = heatmap.collections[0].colorbar
cbar.set_label('Percentage')
plt.show()
# Membuat confusion matrix pada data testing Weighted Ensemble Learning
conf_matrix = confusion_matrix(Y_test, y_pred_weighted)
# Melakukan normalisasi pada confusion matrix dengan satuan persen
conf_matrix_perc = conf_matrix.astype('float') / conf_matrix.sum(axis=1)[:, np.newaxis]
* 100
display_labels = ['Tidak menderita penyakit jantung', 'Menderita Penyakit Jantung']
label_values = [['{:.0f}\n({:.1f}%)'.format(val, perc) for val, perc in zip(row, row_perc)]
for row, row_perc in zip(conf_matrix, conf_matrix_perc)]
```

```
# Membuat heatmap dengan label-label
plt.figure(figsize=(8, 6))
heatmap = sns.heatmap(conf_matrix_perc, annot=label_values, fmt='', cmap='nipy_spectral', cbar=True, xticklabels=display_labels, yticklabels=display_labels)
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix  Weighted Ensemble Learning')
plt.text(1, -0.5, 'Weighted Ensemble Learning Model', fontsize=12, ha='center', va='center', bbox=dict(facecolor='lightgray', edgecolor='gray', boxstyle='round,pad=0.5'))
# Menambahkan color scale legend
cbar = heatmap.collections[0].colorbar
cbar.set_label('Percentage')
plt.show()
#Menghitung nilai ROC AUC
roc_auc_rf = roc_auc_score(Y_test, y_pred_rf)
roc_auc_knn = roc_auc_score(Y_test, y_pred_knn)
roc_auc_weighted = roc_auc_score(Y_test, y_pred_weighted)
#Membuat visualisasi ROC Curve
fpr_rf, tpr_rf, _ = roc_curve(Y_test, y_pred_rf)
fpr_knn, tpr_knn, _ = roc_curve(Y_test, y_pred_knn)
fpr_weighted, tpr_weighted, _ = roc_curve(Y_test, y_pred_weighted)
plt.plot(fpr_rf, tpr_rf, label='Random Forest (AUC = {:.2f})'.format(roc_auc_rf))
plt.plot(fpr_knn, tpr_knn, label='k-NN (AUC = {:.2f})'.format(roc_auc_knn))
plt.plot(fpr_weighted, tpr_weighted, label='Weighted Ensemble Learning (AUC = {:.2f})'.format(roc_auc_weighted))
plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Perbandingan ROC Curve model Random Forest, k-NN,  Weighted Ensemble Learning')
plt.legend(loc='lower right')
```

```
plt.show()
```

C- Lampiran Jumlah Variabel pada *dataset*

Jumlah Variabel Age_Category yang berhubungan dengan Variabel Heart Disease:

	Age_Category	Heart_Disease	Count
0	18-24	No	18587
1	18-24	Yes	94
2	25-29	No	15381
3	25-29	Yes	113
4	30-34	No	18227
5	30-34	Yes	201
6	35-39	No	20332
7	35-39	Yes	274
8	40-44	No	21160
9	40-44	Yes	435
10	45-49	No	20290
11	45-49	Yes	678
12	50-54	No	23916
13	50-54	Yes	1181
14	55-59	No	26063
15	55-59	Yes	1991
16	60-64	No	29406
17	60-64	Yes	3012
18	65-69	No	29611
19	65-69	Yes	3823
20	70-74	No	26542
21	70-74	Yes	4561
22	75-79	No	16953
23	75-79	Yes	3752
24	80+	No	17415
25	80+	Yes	4856

Lampiran 1 Jumlah Variabel *Age Category* yang berhubungan dengan Variabel *Heart Disease*

Jumlah Variabel Height_(cm) yang berhubungan dengan Variabel Heart Disease:

	Height_(cm)	Heart_Disease	Count
0	91.0	No	4
1	91.0	Yes	1
2	94.0	No	4
3	96.0	No	1
4	97.0	No	4
5	97.0	Yes	1
6	99.0	No	2
7	100.0	No	2
8	102.0	No	3
9	103.0	No	1
10	104.0	No	5
11	105.0	No	24
12	105.0	Yes	1
13	106.0	No	3
14	107.0	No	3
15	107.0	Yes	1
16	108.0	No	1
17	110.0	No	1
18	115.0	No	1
19	117.0	No	2
20	119.0	No	2
21	120.0	No	4
22	122.0	No	54
23	122.0	Yes	5
24	124.0	No	14
25	124.0	Yes	1
26	125.0	No	3
27	127.0	No	16
28	127.0	Yes	1
29	130.0	No	21

Lampiran 2 Jumlah Variabel *Height (cm)* yang berhubungan dengan Variabel *Heart Disease*

121	193.0	Yes	239
122	195.0	No	1
123	196.0	No	1263
124	196.0	Yes	86
125	197.0	No	1
126	198.0	No	627
127	198.0	Yes	44
128	200.0	No	4
129	201.0	No	247
130	201.0	Yes	14
131	203.0	No	162
132	203.0	Yes	14
133	205.0	No	2
134	206.0	No	59
135	206.0	Yes	2
136	208.0	No	29
137	208.0	Yes	1
138	211.0	No	22
139	211.0	Yes	2
140	213.0	No	18
141	213.0	Yes	1
142	216.0	No	4
143	216.0	Yes	1
144	218.0	No	5
145	221.0	No	2
146	224.0	No	3
147	226.0	No	7
148	229.0	No	5
149	234.0	No	1
150	241.0	No	1

Lampiran 3 jumlah Variabel *Height (cm)* yang berhubungan dengan Variabel *Heart Disease*

Jumlah Variabel Weight_(kg) yang berhubungan dengan Variabel Heart Disease:

	Weight_(kg)	Heart_Disease	Count
0	24.95	No	1
1	25.40	No	1
2	26.31	No	1
3	26.76	No	1
4	27.22	No	1
5	29.94	No	1
6	30.00	No	1
7	30.00	Yes	1
8	30.84	Yes	1
9	31.75	No	10
10	31.75	Yes	2
11	32.66	No	4
12	32.66	Yes	1
13	33.11	No	1
14	33.57	No	3
15	33.57	Yes	1
16	34.02	No	12
17	34.47	No	3
18	34.93	No	1
19	34.93	Yes	1
20	35.38	No	2
21	35.38	Yes	1
22	35.83	No	2
23	35.83	Yes	1
24	36.29	No	54
25	36.29	Yes	3
26	36.74	No	5
27	36.74	Yes	1
28	37.19	No	13
29	37.19	Yes	2
30	37.65	No	4

Lampiran 4 Jumlah Variabel *Weight (kg)* yang berhubungan dengan Variabel *Heart Disease*

832	226.80	Yes	2
833	227.70	No	2
834	228.16	No	3
835	228.61	No	4
836	229.06	No	2
837	229.52	No	2
838	229.97	No	2
839	230.88	No	3
840	231.33	No	6
841	232.69	No	2
842	233.60	No	3
843	235.87	No	5
844	238.14	No	3
845	238.14	Yes	1
846	240.40	No	2
847	240.40	Yes	1
848	244.03	No	1
849	244.94	No	1
850	247.21	No	1
851	249.48	No	4
852	250.00	No	1
853	252.20	No	1
854	254.01	No	3
855	257.64	No	1
856	258.55	No	1
857	263.08	No	1
858	272.16	No	5
859	272.61	No	1
860	273.52	No	1
861	274.42	No	2
862	283.50	Yes	1
863	285.76	No	1
864	293.02	No	1

Lampiran 5 Jumlah Variabel *Weight (kg)* yang berhubungan dengan Variabel *Heart Disease*

Jumlah Variabel BMI yang berhubungan dengan Variabel Heart Disease:

	BMI	Heart_Disease	Count
0	12.02	No	1
1	12.05	No	1
2	12.11	Yes	1
3	12.12	No	1
4	12.16	No	4
5	12.17	No	1
6	12.20	No	1
7	12.21	No	2
8	12.40	No	1
9	12.48	No	1
10	12.53	No	2
11	12.55	No	2
12	12.55	Yes	1
13	12.65	No	1
14	12.70	No	1
15	12.87	No	1
16	12.88	No	1
17	12.89	No	1
18	12.91	No	6
19	12.92	No	3
20	13.00	No	1
21	13.02	No	1
22	13.05	No	1
23	13.06	No	1
24	13.14	Yes	1
25	13.19	No	1
26	13.20	No	1

Lampiran 6 Jumlah Variabel *BMI* yang berhubungan dengan Variabel *Heart Disease*

5650	83.01	No	1
5651	83.20	No	2
5652	83.26	No	1
5653	83.45	Yes	1
5654	83.68	No	1
5655	84.04	No	2
5656	84.75	No	1
5657	84.87	No	1
5658	85.23	No	1
5659	85.69	Yes	1
5660	85.74	No	1
5661	85.96	No	1
5662	86.32	No	1
5663	86.51	No	3
5664	87.18	Yes	1
5665	87.22	No	1
5666	87.71	No	1
5667	88.57	Yes	1
5668	89.10	No	2
5669	90.39	No	1
5670	91.23	No	2
5671	91.52	No	1
5672	91.82	No	1
5673	92.45	No	1
5674	94.41	No	1
5675	94.94	No	2
5676	96.52	No	1
5677	97.58	No	1
5678	97.65	No	1
5679	98.44	No	1
5680	99.17	No	1
5681	99.33	No	1

Lampiran 7 Jumlah Variabel *BMI* yang berhubungan dengan Variabel *Heart Disease*

Jumlah Variabel Alcohol_Consumption yang berhubungan dengan Variabel Heart Disease:

	Alcohol_Consumption	Heart_Disease	Count
0	0.0	No	125443
1	0.0	Yes	14819
2	1.0	No	23185
3	1.0	Yes	1798
4	2.0	No	18531
5	2.0	Yes	1209
6	3.0	No	10831
7	3.0	Yes	648
8	4.0	No	21980
9	4.0	Yes	1387
10	5.0	No	9165
11	5.0	Yes	457
12	6.0	No	3346
13	6.0	Yes	172
14	7.0	No	2460
15	7.0	Yes	112
16	8.0	No	12825
17	8.0	Yes	613
18	9.0	No	227
19	9.0	Yes	12
20	10.0	No	7501
21	10.0	Yes	380
22	11.0	No	38
23	11.0	Yes	3
24	12.0	No	8374
25	12.0	Yes	451
26	13.0	No	72
27	13.0	Yes	4
28	14.0	No	460
29	14.0	Yes	26
30	15.0	No	5889
31	15.0	Yes	289

Lampiran 8 Jumlah Variabel *Alcohol Consumption* yang berhubungan dengan Variabel *Heart Disease*

Jumlah Variabel Fruit_Consumption yang berhubungan dengan Variabel Heart Disease:

	Fruit_Consumption	Heart_Disease	Count
0	0.0	No	7507
1	0.0	Yes	826
2	1.0	No	2975
3	1.0	Yes	375
4	2.0	No	5192
5	2.0	Yes	600
6	3.0	No	4167
7	3.0	Yes	476
8	4.0	No	12409
9	4.0	Yes	1278
10	5.0	No	4753
11	5.0	Yes	466
12	6.0	No	1495
13	6.0	Yes	191
14	7.0	No	890
15	7.0	Yes	75
16	8.0	No	20407
17	8.0	Yes	1872
18	9.0	No	93
19	9.0	Yes	9
20	10.0	No	4043
21	10.0	Yes	428
22	11.0	No	17
23	11.0	Yes	4
24	12.0	No	27884
25	12.0	Yes	2375
26	13.0	No	32
27	13.0	Yes	4
28	14.0	No	220
29	14.0	Yes	15
30	15.0	No	3695
31	15.0	Yes	371
32	16.0	No	16197

Lampiran 9 *Jumlah* Variabel *fruit consumption* yang berhubungan dengan Variabel *Heart Disease*

Jumlah Variabel Green_Vegetables_Consumption yang berhubungan dengan Variabel Heart Disease:

	Green_Vegetables_Consumption	Heart_Disease	Count
0	0.0	No	19265
1	0.0	Yes	2124
2	1.0	No	7664
3	1.0	Yes	847
4	2.0	No	10293
5	2.0	Yes	1057
6	3.0	No	5899
7	3.0	Yes	562
8	4.0	No	34456
9	4.0	Yes	3253
10	5.0	No	5811
11	5.0	Yes	456
12	6.0	No	2085
13	6.0	Yes	210
14	7.0	No	1369
15	7.0	Yes	115
16	8.0	No	36933
17	8.0	Yes	3396
18	9.0	No	138
19	9.0	Yes	10
20	10.0	No	5781
21	10.0	Yes	473
22	11.0	No	30
23	11.0	Yes	1
24	12.0	No	38704
25	12.0	Yes	3275
26	13.0	No	52
27	13.0	Yes	2
28	14.0	No	354

Lampiran 10 *Jumlah* Variabel *green vegetables consumption* yang berhubungan dengan Variabel *Heart Disease*

Jumlah Variabel FriedPotato_Consumption yang berhubungan dengan Variabel Heart Disease:

	FriedPotato_Consumption	Heart_Disease	Count
0	0.0	No	39723
1	0.0	Yes	4510
2	1.0	No	27325
3	1.0	Yes	2551
4	2.0	No	33289
5	2.0	Yes	2806
6	3.0	No	14190
7	3.0	Yes	1157
8	4.0	No	62627
9	4.0	Yes	5206
10	5.0	No	9440
11	5.0	Yes	681
12	6.0	No	2775
13	6.0	Yes	236
14	7.0	No	1752
15	7.0	Yes	132
16	8.0	No	35414
17	8.0	Yes	2952
18	9.0	No	129
19	9.0	Yes	16
20	10.0	No	7041
21	10.0	Yes	508
22	11.0	No	38
23	11.0	Yes	2
24	12.0	No	20285
25	12.0	Yes	1655
26	13.0	No	47
27	13.0	Yes	3
28	14.0	No	246
29	14.0	Yes	16

Lampiran 11 *Jumlah* Variabel *fried potato consumption* yang berhubungan dengan Variabel *Heart Disease*

```
Jumlah nilai pada variabel 'Height_(cm)':
168.0    27119
163.0    25804
170.0    24739
178.0    24673
173.0    23591
165.0    23551
175.0    22059
183.0    22008
160.0    20829
157.0    19377
180.0    19151
185.0    11081
155.0     9897
152.0     9250
188.0     9094
191.0     5112
193.0     3170
150.0     2918
196.0     1349
147.0      994
198.0      671
145.0      528
142.0      263
201.0      261
203.0      176
140.0      113
```

Lampiran 12 Jumlah Variabel *Height (cm)*

```
Jumlah nilai pada variabel 'Weight_(kg)':
90.72     16614
81.65     15104
72.57     13263
68.04     13240
77.11     12216
86.18     11083
63.50      9583
79.38      9066
74.84      8559
99.79      8454
83.91      7970
95.25      7590
58.97      7279
65.77      6651
104.33     6628
113.40     6377
70.31      6369
61.23      6247
108.86     5517
88.45      5452
54.43      4563
97.52      4353
56.70      4313
102.06     4026
92.99      3930
117.93     3218
```

Lampiran 13 Jumlah Variabel *Weight (kg)*

```
Jumlah nilai pada variabel 'BMI':
26.63    3340
27.46    2658
24.41    2596
27.44    2567
27.12    2259
25.10    2068
32.28    1935
28.70    1894
29.53    1881
29.29    1853
25.84    1831
25.09    1657
28.34    1609
24.21    1592
29.05    1559
25.75    1557
25.06    1555
26.61    1548
24.96    1510
26.58    1479
28.89    1473
25.83    1470
28.19    1466
```

Lampiran 14 Jumlah Variabel *BMI*

```
Jumlah nilai pada variabel 'Alcohol_Consumption':
0.0     140262
1.0      24983
4.0      23367
2.0      19740
8.0      13438
30.0     11976
3.0      11479
5.0       9622
20.0      9372
12.0      8825
10.0      7881
15.0      6178
28.0      5935
16.0      3602
6.0       3518
25.0      2670
7.0       2572
24.0      1253
14.0       486
29.0       309
9.0        239
```

Lampiran 15 Jumlah Variabel *Alcohol Consumption*

```
Jumlah nilai pada variabel 'Fruit_Consumption':
30.0     90273
60.0     48127
12.0     30259
8.0      22279
20.0     17476
16.0     17444
90.0     16567
4.0      13687
0.0       8333
2.0       5792
5.0       5219
3.0       4643
10.0      4471
15.0      4066
28.0      3379
1.0       3350
120.0     3113
24.0      2409
40.0      2003
6.0       1686
25.0      1137
7.0        965
32.0       322
14.0       235
99.0       199
48.0       145
45.0       115
56.0       108
```

Lampiran 16 Jumlah Variabel *Fruit Consumption*

```
Jumlah nilai pada variabel 'Green_Vegetables_Consumption':
30.0     51162
12.0     41979
8.0      40329
4.0      37709
16.0     23338
0.0      21389
20.0     21288
2.0      11350
60.0      8796
1.0       8511
15.0      7002
3.0       6461
5.0       6267
10.0      6254
24.0      2788
28.0      2678
6.0       2295
25.0      2204
90.0      1944
7.0       1484
40.0       955
120.0      615
14.0       375
18.0       170
99.0       168
9.0        148
```

Lampiran 17 Jumlah Variabel *Green Vegetables Consumption*

```
Jumlah nilai pada variabel 'FriedPotato_Consumption':
4.0      67833
0.0      44233
8.0      38366
2.0      36095
1.0      29876
12.0     21940
3.0      15347
5.0      10121
30.0      8434
16.0      8038
10.0      7549
20.0      6908
15.0      4559
6.0       3011
7.0       1884
60.0      1082
24.0       579
28.0       555
25.0       522
90.0       461
14.0       262
40.0       234
120.0      205
9.0        145
18.0        96
13.0        50
17.0        49
```

Lampiran 18 Jumlah Variabel *Fried Potato Consumption*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 308854 entries, 0 to 308853
Data columns (total 19 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   General_Health                308854 non-null  object
 1   Checkup                       308854 non-null  object
 2   Exercise                      308854 non-null  object
 3   Heart_Disease                 308854 non-null  object
 4   Skin_Cancer                   308854 non-null  object
 5   Other_Cancer                  308854 non-null  object
 6   Depression                    308854 non-null  object
 7   Diabetes                      308854 non-null  object
 8   Arthritis                     308854 non-null  object
 9   Sex                           308854 non-null  object
 10  Age_Category                  308854 non-null  object
 11  Height_(cm)                   308854 non-null  float64
 12  Weight_(kg)                   308854 non-null  float64
 13  BMI                           308854 non-null  float64
 14  Smoking_History               308854 non-null  object
 15  Alcohol_Consumption           308854 non-null  float64
 16  Fruit_Consumption             308854 non-null  float64
 17  Green_Vegetables_Consumption  308854 non-null  float64
 18  FriedPotato_Consumption       308854 non-null  float64
dtypes: float64(7), object(12)
memory usage: 44.8+ MB
```

Lampiran 19 Hasil Pengumpulan informasi data

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_(cm)	Weight_(kg)	BMI	Smoking_History
0	3	2	0	0	0	0	0	0	1	0	10	150.0	32.66	14.54	1
1	4	4	0	1	0	0	0	2	0	0	10	165.0	77.11	28.29	0
2	4	4	1	0	0	0	0	2	0	0	8	163.0	88.45	33.47	0
3	3	4	1	1	0	0	0	2	0	1	11	180.0	93.44	28.73	0
4	2	4	0	0	0	0	0	0	0	1	12	191.0	88.45	24.37	1

Lampiran 20 Hasil *Encoding Data*

Arthritis	Sex	Age_Category	Height_(cm)	Weight_(kg)	BMI	Smoking_History	Alcohol_Consumption	Fruit_Consumption	Green_Vegetables_Consumption	FriedPotato_Consumption
1	0	10	150.0	32.66	14.54	1	0.0	30.0	16.0	12.0
0	0	10	165.0	77.11	28.29	0	0.0	30.0	0.0	4.0
0	0	8	163.0	88.45	33.47	0	4.0	12.0	3.0	16.0
0	1	11	180.0	93.44	28.73	0	0.0	30.0	30.0	8.0
0	1	12	191.0	88.45	24.37	1	0.0	8.0	4.0	0.0

Lampiran 21 Hasil *Encoding Data*

```
General_Health                0
Checkup                       0
Exercise                      0
Heart_Disease                 0
Skin_Cancer                   0
Other_Cancer                  0
Depression                    0
Diabetes                      0
Arthritis                     0
Sex                           0
Age_Category                  0
Height_(cm)                   0
Weight_(kg)                   0
BMI                           0
Smoking_History               0
Alcohol_Consumption           0
Fruit_Consumption             0
Green_Vegetables_Consumption  0
FriedPotato_Consumption       0
dtype: int64
```

Lampiran 22 Hasil pengumpulan pemeriksaan data

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_(cm)	Weight_(kg)	BMI	Smoking_History	
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
308849	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
308850	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
308851	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
308852	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
308853	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals

308854 rows × 19 columns

Lampiran 23 Hasil pengumpulan pemeriksaan data

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	\
46402	Good	Within the past year	Yes	No	No	No	Yes	No	No	Female	
49287	Very Good	Within the past year	Yes	No	No	No	No	No	No	Female	
75448	Excellent	Within the past year	Yes	No	No	No	No	No	No	Female	
76857	Excellent	Within the past year	Yes	No	No	No	No	No	No	Male	
78871	Good	Within the past year	Yes	No	No	No	No	No	No	Female	
...	...	...	...	...	...	...	...	...	...	...	
301474	Good	Within the past year	Yes	No	No	No	No	No	Yes	Female	
303040	Very Good	Within the past year	Yes	No	No	No	No	No	No	Female	
303600	Good	Within the past year	Yes	No	No	No	No	No	No	Female	
303609	Very Good	Within the past year	Yes	No	No	No	No	No	No	Female	
308375	Very Good	Within the past year	Yes	No	No	No	No	No	No	Male	

Lampiran 24 Hasil data yang terduplikasi

```
[[ 2.  0.  0. ...  0. 30. 16.]
 [ 4.  0.  1. ...  0. 30.  0.]
 [ 4.  1.  0. ...  4. 12.  3.]
 ...
 [ 0.  1.  0. ...  4. 40.  8.]
 [ 4.  1.  0. ...  3. 30. 12.]
 [ 4.  1.  0. ...  1.  5. 12.]]
```

Lampiran 25 Variabel X dengan *standard scaler*

```
[[12.]
 [ 4.]
 [16.]
 ...
 [ 4.]
 [ 0.]
 [ 1.]]
```

Lampiran 26 Variabel Y dengan *standard scaler*