Memprediksi Risiko Serangan Jantung dari Faktor Pemodelan Data Kesehatan Melalui GaussianNB dan Logistic Regression

Predicting Heart Attack Risk from Health Data Modeling Factors through GaussianNB and Logistic Regression

Jonathan Chandra[1]
[1] Fakultas Teknik Informasi
Universitas Multimedia Nusantara
Serpong, Indonesia
[1] jonathan.chandra@student.umn.ac.id

***Abstract*—Serangan jantung adalah salah satu penyebab utama kematian di seluruh dunia. Penelitian ini bertujuan untuk menganalisis hubungan antara gaya hidup dan faktor kesehatan dengan risiko serangan jantung dengan menggunakan pendekatan *machine learning*. Dua algoritma, Naïve Bayes dan *Logistic Regression*, digunakan untuk memprediksi risiko serangan jantung berdasarkan data yang mencakup variabel seperti kolesterol, tekanan darah, BMI, kebiasaan merokok, dan tingkat stres. Hasil evaluasi menunjukkan bahwa *Logistic Regression* memiliki akurasi yang lebih tinggi (70%) dibandingkan dengan Naïve Bayes (67%), meskipun perbedaannya tidak signifikan. Analisis korelasi menunjukkan bahwa faktor-faktor seperti kolesterol, obesitas, merokok, dan diabetes memiliki hubungan positif yang signifikan dengan risiko serangan jantung, sedangkan olahraga memiliki hubungan negatif. Penelitian ini menegaskan pentingnya manajemen gaya hidup dalam mengurangi risiko serangan jantung. Selain itu, hasil penelitian ini menunjukkan potensi pembelajaran mesin sebagai alat prediksi yang efektif dalam mendukung pengambilan keputusan berbasis data untuk intervensi kesehatan masyarakat.**

Keywords—Machine Learning, Logistic Regression, Naïve Bayes, Risiko Serangan Jantung, Gaya Hidup, Serangan Jantung.

I. Introduction

Penyakit jantung merupakan salah satu tantangan kesehatan utama yang menyebabkan tingginya angka kematian di seluruh dunia, menyebabkan sekitar 17,8 juta kematian setiap tahun menurut laporan WHO 2023, yang mencangkup 32% dari total kematian global [2]. Data terbaru menunjukkan bahwa penyakit ini merupakan penyebab utama kematian di dunia, menyumbang hampir sepertiga dari semua kematian. Upaya pencegahan dini menjadi fokus utama berbagai pihak, termasuk rumah sakit dan pemerintah, untuk mengurangi dampak ekonomi dari mahalnya biaya pengobatan [1]. Kondisi ini mendorong perlunya upaya pencegahan yang lebih efektif, salah satunya melalui pemanfaatan teknologi analisis data. Model prediktif berbasis kecerdasan buatan dan algoritma machine learning menjadi alternatif strategis dalam mendeteksi risiko penyakit jantung sejak dini [3]. Algoritma seperti *Gaussian Naïve Bayes* (GNB) dan *Logistic Regression* telah terbukti dapat menangani data kesehatan dengan baik untuk keperluan klasifikasi dan prediksi. GNB, dengan asumsi independensi antar variabel, menawarkan kecepatan pemrosesan yang tinggi, sementara *Logistic Regression* memberikan hasil yang dapat diinterpretasikan dan efektif dalam memodelkan hubungan antar variabel [5]. Penelitian ini bertujuan untuk mengeksplorasi keandalan kedua metode tersebut dalam mengidentifikasi faktor risiko serangan jantung, sehingga mendukung tindakan pencegahan yang lebih akurat dan berkelanjutan.

Penerapan *machine learning* dalam deteksi dini berbagai penyakit semakin banyak digunakan karena mampu membuat kesimpulan yang sejalan dengan informasi baru dengan mendeteksi pola-pola tersembunyi dari sebuah *dataset* [6]. Machine Learning dapat memberikan kontribusi yang signifikan terhadap dunia medis dengan memberikan diagnosis penyakit yang akurat dan efisien [4]. Karena deteksi dini penyakit jantung dapat menurunkan angka kematian, sehingga salah satu cara efektif dalam mengidentifikasi dan memprediksi penyakit jantung adalah dengan menggunakan untuk *algoritma machine learning* [7]. *Machine Learning* dapat membuat model yang dapat digunakan untuk membuat prediksi atau keputusan berdasarkan data hasil observasi [10]. *Machine Learning* mampu mengatasi berbagai kerumitan dalam proses mendiagnosis penyakit jantung dengan model mengatasi berbagai kerumitan dalam proses mendiagnosis penyakit jantung dengan model prediksi menggunakan beberapa algoritma seperti *Gaussian Naive Bayes* (GNB), *Logistic Regression* (LR), *support vector machine* (SVM), *k-nearest neighbor algorithm* (k-NN), *artificial neural network* (ANN), *decision tree* (DT), *AdaBoost* (AB), *logika fuzzy* (FL), *Extreme Gradient Boosting* (XGBoost), dan lainnya, Tetapi untuk penelitian ini menggunakan *Gaussian Naive Bayes* (GaussianNB) dan *Logistic Regression* (LR) [4] [5].

Salah satu faktor yang menyebabkan terjadinya serangan jantung berasal dari kondisi dimana aliran darah yang membawa oksigen ke otot jantung tiba-tiba yang membawa oksigen ke otot jantung tiba - tiba tersumbat. Sehingga jantung tidak akan mendapatkan cukup oksigen yang dapat berakibat pada rusaknya otot jantung. Penyebab umum penyumbatan mendadak dikarenakan pembentukan bekuan darah (thrombus). Bekuan darah terbentuk akibat adanya aterosklerosis, suatu kondisi di mana timbunan lemak (plak) menumpuk di dinding bagian dalam pembuluh darah [8]. Sebagian besar serangan jantung disebabkan oleh atherosclerosis. Namun, faktor usia, gaya hidup yang tidak sehat, serta kondisi medis lainnya dapat meningkatkan resiko terkena serangan jantung [12] [14].

Metode pembelajaran mesin Regresi Logistik dan *Gaussian Naive Bayes* (GNB) dapat digunakan untuk memprediksi variabel target Risiko Serangan Jantung

berdasarkan kumpulan data yang mencakup faktor-faktor seperti usia, jenis kelamin, kolesterol, tekanan darah, detak jantung, riwayat keluarga, kebiasaan merokok, dan tingkat aktivitas fisik [11]. Regresi Logistik memodelkan probabilitas risiko serangan jantung dengan menganalisis hubungan linier antara faktor-faktor risiko tersebut dan target, menghasilkan nilai probabilitas yang diklasifikasikan sebagai risiko rendah atau tinggi [13]. Sebaliknya, GNB menghitung probabilitas risiko serangan jantung menggunakan *Teorema Bayes* dengan asumsi bahwa fitur-fitur seperti BMI, trigliserida, dan konsumsi alkohol mengikuti distribusi *Gaussian Naïve Bayes* [12]. Kedua metode ini bekerja dengan memanfaatkan pola dalam data untuk memprediksi kemungkinan risiko serangan jantung secara akurat, sehingga memungkinkan pengambilan keputusan medis yang lebih baik berdasarkan karakteristik masing-masing pasien [14].

II. Study Literature

2.1 Machine Learning

Machine Learning (ML) adalah cabang dari kecerdasan buatan yang memungkinkan sistem komputer untuk mempelajari data dan melakukan tugas-tugas tertentu tanpa memerlukan pemrograman eksplisit. Menurut Mahesh B. (2020), ML melibatkan pengembangan algoritma dan model yang dapat belajar dari data dan meningkatkan kinerja dari waktu ke waktu melalui pengalaman [16]. Dalam konteks analisis data, ML digunakan sebagai metode untuk menemukan pola, hubungan, dan wawasan yang tersembunyi dari kumpulan data yang besar dan kompleks. Dengan memanfaatkan algoritma seperti supervised learning, unsupervised learning, dan reinforcement learning, ML dapat memproses data secara efisien, mengenali pola yang tidak terlihat secara langsung, dan memberikan prediksi yang akurat. Sebagai contoh, dalam analisis data medis, ML dapat mengidentifikasi risiko penyakit berdasarkan data pasien, seperti tekanan darah, kadar kolesterol, dan riwayat keluarga [15]. Teknik ini sangat berharga karena menyediakan pendekatan berbasis data yang dapat mendukung pengambilan keputusan yang lebih tepat di berbagai bidang seperti keuangan, pemasaran, dan perawatan kesehatan. Hal ini menjadikan ML sebagai alat analisis data yang inovatif dan terus berkembang untuk menjawab tantangan data modern [17].

2.2 Gaussian Naïve Bayes (GNB)

Metode *Gaussian Naïve Bayes* (GNB) adalah algoritma pembelajaran mesin berbasis probabilitas yang digunakan untuk klasifikasi, dengan asumsi bahwa setiap fitur input mengikuti distribusi Gaussian (normal). GNB didasarkan pada Teorema Bayes, yang menggabungkan probabilitas prior dan likelihood untuk menghitung probabilitas posterior, yaitu kemungkinan kelas berdasarkan data yang diamati [19]. Algoritma ini mengasumsikan independensi antar fitur, yang berarti bahwa nilai dari satu fitur tidak mempengaruhi nilai dari fitur lain dalam menentukan hasil klasifikasi. Pendekatan ini membuat GNB menjadi sangat efisien, terutama untuk dataset dengan banyak fitur, karena hanya perlu menghitung parameter mean dan varians untuk setiap fitur. GNB sering digunakan dalam aplikasi seperti klasifikasi teks, deteksi spam, diagnosis medis, dan analisis risiko karena kesederhanaan, kecepatan, dan kemampuannya untuk bekerja dengan dataset yang kecil. Meskipun asumsi independensi sering kali tidak realistis, GNB masih memberikan kinerja yang baik dalam praktiknya, menjadikannya metode yang dapat diandalkan untuk banyak masalah klasifikasi [20].

2.3 Logistic Regression (LR)

Logistic Regression adalah metode machine learning yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen dan variabel dependen biner (dua kelas). Metode ini bekerja dengan memanfaatkan fungsi logistik (sigmoid) untuk memetakan output yang diprediksi dalam rentang probabilitas antara 0 dan 1, yang kemudian digunakan untuk menentukan kelas output berdasarkan ambang batas tertentu, biasanya 0,5. Regresi Logistik mengasumsikan hubungan linier antara variabel independen dan log odds (logaritma peluang), sehingga model ini cocok untuk data dengan hubungan linier [18]. Parameter model diestimasi dengan menggunakan metode estimasi kemungkinan maksimum (maximum likelihood estimation, MLE), yang memaksimalkan kemungkinan hasil prediksi sesuai dengan data aktual. *Logistic Regression* banyak digunakan dalam berbagai aplikasi seperti diagnosis medis, analisis risiko keuangan, dan prediksi churn pelanggan karena kesederhanaannya, kemampuan interpretasi, dan efisiensi komputasi. Meskipun terbatas pada hubungan linier, metode ini dapat diperluas ke masalah klasifikasi multi kelas menggunakan pendekatan seperti regresi one-vs-rest atau softmax [21].

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X$$

Ketika menyelesaikan persamaan ini untuk probabilitas (P), probabilitas memiliki hubungan sigmoidal dengan variabel independen (Gambar A), dan estimasi probabilitas sekarang dibatasi dengan tepat antara 0 dan 1. Tidak seperti regresi linier yang memodelkan hubungan linier antara variabel independen dan hasil, Regresi Logistik memodelkan hubungan linier antara logit (logaritma natural peluang) dari hasil dan variabel independen. Koefisien regresi dalam model ini mewakili intersep (b0) dan gradien (b1) dari hubungan linier [23]. Untuk memastikan probabilitas yang diprediksi berada dalam kisaran 0 hingga 1, Regresi Logistik menggunakan fungsi logistik atau sigmoid yang menghasilkan hubungan berbentuk sigmoid antara variabel independen dan probabilitas hasil [22].

Regresi Logistik juga dapat diperluas untuk mengakomodasi lebih dari satu variabel independen. Hal ini memungkinkan peneliti untuk menganalisis hubungan setiap variabel dengan hasil biner sambil menjaga nilai variabel independen lainnya tetap konstan. Pendekatan ini sangat berguna untuk memahami hubungan independen antara variabel dan hasil tertentu dan untuk mengendalikan pengaruh variabel perancu dalam studi observasi. Model ini fleksibel dan banyak digunakan dalam berbagai aplikasi seperti prediksi risiko, diagnosis medis, dan analisis sosial [24].

III. Methodology

3.1 CRISP - DM

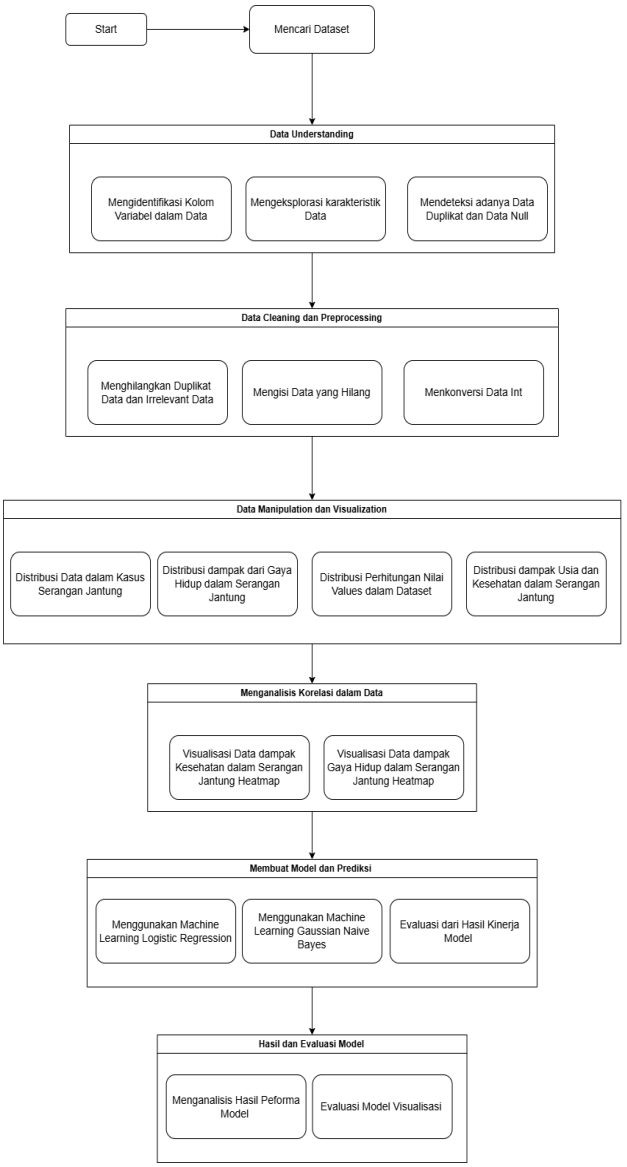
CRISP-DM *(Cross-Industry Standard Process for Data Mining)* adalah metodologi standar yang sering digunakan dalam proyek-proyek data mining dan analisis data, karena

fleksibilitas dan kemampuannya untuk diterapkan di berbagai industri. Metodologi ini terdiri dari enam langkah utama yang mencakup seluruh siklus proyek: Pemahaman Bisnis, Pemahaman Data, Persiapan Data, Pemodelan, Evaluasi, dan Penerapan. Langkah pertama, Pemahaman Bisnis, bertujuan untuk memahami masalah bisnis dan merumuskan tujuan analisis yang jelas [25]. Pemahaman Data melibatkan pengumpulan, eksplorasi, dan verifikasi kualitas data untuk memastikan relevansi dan keandalannya. Pada tahap Data Preparation, data diproses melalui seleksi, pembersihan, transformasi, dan penggabungan untuk siap digunakan dalam pemodelan [27] [28]. Pemodelan kemudian dilakukan dengan memilih algoritma yang sesuai, melatih model, dan mengevaluasi kinerjanya pada tahap awal. Tahap Evaluasi menilai hasil model dengan menggunakan matrik tertentu untuk memastikan relevansi dengan tujuan bisnis dan mengidentifikasi kesalahan atau bias. Terakhir, Deployment melibatkan penerapan model ke dalam sistem operasional, menyajikan hasil untuk mendukung pengambilan keputusan, dan memelihara model agar tetap relevan [29]. CRISP-DM merupakan proses yang berulang, yang memungkinkan langkah-langkah sebelumnya diulang jika diperlukan untuk memastikan keakuratan dan kesesuaian dengan kebutuhan proyek, menjadikannya kerangka kerja yang sistematis dan efektif untuk manajemen proyek penggalian data [26].

3.2 Proses Flow

Proses alur kerja berbasis *Cross-Industry Standard Process for Data Mining* (CRISP-DM) dalam visualisasi ini dimulai dengan tahap *Data Understanding*, yang melibatkan pengumpulan dataset, mengidentifikasi bidang variabel, mengeksplorasi karakteristik data, dan mendeteksi nilai duplikat dan *missing value* [26]. Tahap selanjutnya adalah *Data Cleaning* dan *Preprocessing*, di mana data yang tidak relevan atau duplikat dihapus, nilai yang hilang diisi, dan data dikonversi ke format yang sesuai seperti *integer*. Setelah dibersihkan, tahap *Data Manipulation* dan *Visualization* dilakukan untuk menganalisis distribusi data, dampak gaya hidup terhadap kasus serangan jantung, dan nilai variabel lainnya, yang divisualisasikan untuk memahami pola dasar. Tahap selanjutnya adalah *Correlation Analysis*, dimana hubungan antar fitur diuji dengan menggunakan *heatmap* untuk mengevaluasi faktor-faktor yang signifikan terhadap serangan jantung.

Proses berlanjut ke tahap *Model Building* and *Prediction*, di mana berbagai algoritma machine learning seperti *Logistic Regression* dan *Gaussian Naïve Bayes* digunakan untuk membuat prediksi berdasarkan data yang telah disiapkan. Evaluasi model dilakukan dengan mengukur kinerja model menggunakan metrik yang relevan. Terakhir, hasil model dianalisis dalam tahap *Evaluasi Model*, yang meliputi analisis kinerja prediksi dan visualisasi hasil evaluasi untuk memverifikasi bahwa model yang dibuat telah memenuhi tujuan analisis awal. Alur proses ini memastikan pendekatan yang sistematis untuk menghasilkan wawasan dan prediksi yang akurat dari data.



Gambar 1. Process Flow Penelitian

3.3 Pencarian dan Import Dataset

Dataset yang digunakan dalam penelitian ini adalah Heart Attack Risk Prediction Dataset, sebuah dataset sintetis yang dirancang untuk mengeksplorasi berbagai faktor yang mempengaruhi kesehatan jantung dan risiko serangan jantung. Dataset ini mencakup 8763 catatan pasien dari berbagai belahan dunia, dengan berbagai atribut yang relevan, termasuk karakteristik individu seperti usia, jenis kelamin, kadar kolesterol, tekanan darah, dan detak jantung. Selain itu, faktor gaya hidup seperti kebiasaan merokok, konsumsi alkohol, pola olahraga, kebiasaan makan, tingkat stres, dan waktu duduk juga dipertimbangkan. Atribut medis seperti riwayat penyakit jantung, penggunaan obat, dan kadar trigliserida juga disertakan, serta aspek sosio ekonomi seperti pendapatan dan lokasi geografis. Data ini berakhir dengan fitur klasifikasi biner yang mengindikasikan risiko serangan jantung, sehingga menjadi sumber daya yang berharga untuk analisis prediktif dan penelitian dalam meningkatkan pemahaman dan pencegahan penyakit kardiovaskular.

Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	Alcohol Consumption	Exercise Hours Per Week
HQM9364	75	Female	136	141/85	101	No	No	0	1	Light	14.744881
ESJ9954	62	Male	262	137/82	89	Yes	No	0	1	Light	16.228489
ONA1218	72	Female	126	138/93	86	Yes	Yes	1	0	None	6.818887
UBE5339	18	Female	300	132/94	109	Yes	No	1	1	Light	18.297860
LUQ7367	67	Female	223	91/89	84	Yes	Yes	1	0	Moderate	10.980701

Previous Heart Problems	Medication Use	Stress Level	Sedentary Hours Per Day	Income	BMI	Triglycerides	Physical Activity Days Per Week	Sleep Hours Per Day	Country	Continent	Hemisp
0	1	4	10.922177	94152	30.589796	374	3	4	Nigeria	Africa	Nort Hemisp
0	1	7	1.208610	159792	31.584511	678	0	8	Australia	Australia	Sout Hemisp
0	1	4	9.514556	254952	34.711478	736	1	5	Argentina	South America	Sout Hemisp
1	1	6	9.015221	25229	29.022289	152	2	5	Spain	Europe	Sout Hemisp
0	1	4	10.020410	229179	35.966244	744	5	8	Japan	Asia	Nort Hemisp

Gambar 2. Gambar Dataset Kolom

3.4 Data Understanding

Pada langkah ini melakukan sebuah eksplorasi data menggunakan *Google Collab* dengan menggunakan *library pandas* dalam proses ini dilakukan untuk membaca dataset untuk melakukan sebuah pengolahan data yang dilakukan [30]. Mengidentifikasi Duplikat Data dan Irrelevant Data yang dilakukan untuk sebuah Data Preparation untuk memahami struktur, karakteristik, dan kualitas dataset yang digunakan. Pada tahap ini memulai untuk mengenali pola, distribusi, dan hubungan antar variabel yang ada dalam dataset tersebut.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8763 entries, 0 to 8762
Data columns (total 26 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Patient ID                      8763 non-null   object
 1   Age                             8763 non-null   int64
 2   Sex                             8763 non-null   object
 3   Cholesterol                     8763 non-null   int64
 4   Blood Pressure                  8763 non-null   object
 5   Heart Rate                      8763 non-null   int64
 6   Diabetes                        8763 non-null   int64
 7   Family History                  8763 non-null   int64
 8   Smoking                         8763 non-null   int64
 9   Obesity                         8763 non-null   int64
 10  Alcohol Consumption             8763 non-null   int64
 11  Exercise Hours Per Week         8763 non-null   float64
 12  Diet                            8763 non-null   object
 13  Previous Heart Problems         8763 non-null   int64
 14  Medication Use                  8763 non-null   int64
 15  Stress Level                    8763 non-null   int64
 16  Sedentary Hours Per Day         8763 non-null   float64
 17  Income                          8763 non-null   int64
 18  BMI                             8763 non-null   float64
 19  Triglycerides                   8763 non-null   int64
 20  Physical Activity Days Per Week 8763 non-null   int64
 21  Sleep Hours Per Day             8763 non-null   int64
 22  Country                         8763 non-null   object
 23  Continent                       8763 non-null   object
 24  Hemisphere                      8763 non-null   object
 25  Heart Attack Risk               8763 non-null   int64
dtypes: float64(3), int64(16), object(7)
```

Gambar 3. Dataset Info

Berdasarkan informasi dari DataFrame, set data ini terdiri dari 8.763 entri, yang mencakup 26 kolom dengan berbagai tipe data. Setiap kolom memiliki jumlah nilai yang sama (8.763), yang mengindikasikan bahwa tidak ada nilai yang hilang dalam dataset. Kolom-kolom ini mencakup informasi tentang pasien, seperti ID Pasien *(objek)*, Usia *(objek)*, Jenis Kelamin *(objek)*, serta berbagai atribut kesehatan seperti Kolesterol, Tekanan Darah, Denyut Jantung, dan Diabetes (semuanya bertipe *int64*).

Selain itu, terdapat beberapa kolom yang menggambarkan kebiasaan atau gaya hidup, seperti Merokok, Obesitas, Konsumsi Alkohol, dan Jam Olahraga Per Minggu. Kolom-kolom ini memiliki tipe data numerik (*int64* atau *float64*). Dataset ini juga mencakup kolom-kolom seperti Pendapatan, BMI, Trigliserida, dan Risiko Serangan Jantung, yang penting untuk analisis risiko kardiovaskular. Beberapa bidang geografis, seperti Negara, Benua, dan Belahan Bumi, merupakan tipe objek, sementara yang lain termasuk atribut numerik seperti Hari Aktivitas Fisik Per Minggu dan Jam Tidur Per Hari. Informasi ini memberikan dasar yang kuat untuk analisis lebih lanjut mengenai kesehatan, kebiasaan, dan faktor risiko serangan jantung di berbagai kelompok populasi.

Hasil *Data Understanding* dalam penelitian ini menunjukkan adanya tidak ada terlihat data duplikat yang ditemukan dalam hasil dalam pengecekkan data duplikat menggunakan Kode *df.duplicate().sum()* dalam analisis kami untuk menghitung jumlah baris duplikat dalam sebuah Dataframe (df). Fungsi *duplicated ()* akan mengembalikan nilai boolean *(True or False)* untuk setiap baris, dimana nilai *True* menunjukkan bahwa garis tersebut adalah duplikat dari baris sebelumnya. Kemudian, fungsi *sum()* akan menghitung total jumlah nilai *True,* yang mewakili jumlah baris duplikat dalam *DataFrame.* Hasil akhir dari eksekusi kode ini adalah sebuah angka yang menunjukkan berapa banyak baris duplikat yang ada dalam *Data Frame* tersebut.

```
# Checking for duplicate row

df.duplicated().sum()
```

```
0
```

Gambar 4. Dataset Checking Data Duplikat

```
Missing Data Summary:
                                Missing Values  Missing Percentage
Patient ID                                   0                 0.0
Age                                          0                 0.0
Sex                                          0                 0.0
Cholesterol                                  0                 0.0
Blood Pressure                               0                 0.0
Heart Rate                                   0                 0.0
Diabetes                                     0                 0.0
Family History                               0                 0.0
Smoking                                      0                 0.0
Obesity                                      0                 0.0
Alcohol Consumption                          0                 0.0
Exercise Hours Per Week                      0                 0.0
Diet                                         0                 0.0
Previous Heart Problems                      0                 0.0
Medication Use                               0                 0.0
Stress Level                                 0                 0.0
Sedentary Hours Per Day                      0                 0.0
Income                                       0                 0.0
BMI                                          0                 0.0
Triglycerides                                0                 0.0
Physical Activity Days Per Week              0                 0.0
Sleep Hours Per Day                          0                 0.0
Country                                      0                 0.0
Continent                                    0                 0.0
Hemisphere                                   0                 0.0
Heart Attack Risk                            0                 0.0
```

Gambar 5. Dataset Checking *Data Null*

Gambar di atas memberikan sebuah ringkasan data yang menunjukkan bahwa tidak ada *missing value* pada pada setiap kolom dalam dataset. Kolom-kolom tersebut mencakup berbagai variabel yang berkaitan dengan kesehatan pasien, seperti ID pasien, usia, jenis kelamin, kolesterol, tekanan darah, dan risiko serangan jantung. Dengan tidak adanya nilai yang hilang, hal ini mengindikasikan bahwa data yang digunakan untuk analisis atau pemodelan cukup lengkap dan bersih. Tidak adanya missing value sangat penting karena akan menghindari bias dalam analisis dan meningkatkan keakuratan hasil yang diperoleh.

3.5 Data Cleaning and Preprocessing

```
df=df.rename(columns={
    "Sleep Hours Per Day":"Sleeping Hours",
    "Physical Activity Days Per Week":"Activity Per Week",
    "Sedentary Hours Per Day":"Sedentary Hours",
    "Patient ID":"ID"

})
```

Gambar 6. Mengubah Nama Kolom Dataset

Kode *df = df.rename(columns = {})* digunakan untuk mengganti nama kolom dalam *Data Frame* (df) di Python, khususnya di perpustakaan Pandas. Pada contoh ini, nama kolom *"Sleep Hours Per Day"* diubah menjadi *"Sleeping Hours"*, *"Physical Activity Days Per Week"* menjadi *"Activity Per Week"*, *"Sedentary Hours Per Day"* menjadi *"Sedentary Hours"*, dan *"Patient ID"* menjadi "ID". Dengan kata lain, kode ini membuat nama-nama kolom menjadi lebih pendek dan mudah dipahami, sehingga memudahkan analisis data lebih lanjut.

Kemudian dalam melakukan *Data Cleaning* akan melakukan drop dalam kolom variabel "ID" karena tipe data tersebut kategorik tidak bagus untuk melakukan Data visualisasi. Selain itu Kode *df["Active Hours"]=(24-df["Sedentary Hours"])* digunakan untuk membuat kolom baru yang disebut *"Active Hours"* dalam *Data Frame* (df) di Python. Kolom baru ini berisi nilai yang dihitung dengan mengurangkan nilai di kolom *"Sedentary Hours"* dari 24. 24 diasumsikan sebagai total jam dalam sehari. Dengan kata lain, kolom *"Active Hours"* akan menunjukkan perkiraan jumlah jam yang tidak dihabiskan untuk aktivitas yang tidak banyak bergerak dalam sehari. Operasi ini sering dilakukan dalam analisis data kesehatan atau gaya hidup untuk mendapatkan pemahaman yang lebih baik tentang aktivitas fisik seseorang.

Melakukan sebuah pengelompokkan Data berdasarkan kriteria kesehatan tertentu, kemudian membuat dua kolom baru yaitu "Health Risk Score" dan "Lifestyle Risk Score". Kolom "Health Risk Score" dihitung berdasarkan beberapa faktor risiko seperti kadar kolesterol, riwayat diabetes, merokok, indeks massa tubuh (BMI), dan tingkat stres. Sementara itu, kolom "Lifestyle Risk Score" dihitung berdasarkan faktor-faktor gaya hidup seperti merokok, obesitas, konsumsi alkohol, durasi aktivitas menetap, dan pola makan. Setiap faktor risiko yang terpenuhi akan memberikan nilai 1, dan kemudian nilai-nilai tersebut dijumlahkan untuk mendapatkan skor risiko total. Dengan cara ini, data dapat dikelompokkan menjadi beberapa kategori risiko berdasarkan skor yang diperoleh, sehingga memudahkan dalam analisis lebih lanjut.

Selanjutnya membuat sebuah kode yang memisahkan nilai tekanan darah atau "*Blood Pressure*" karena berisi 1 kolom dan membuat pemisahan dalam kolom tersebut menjadi *"systolic"* dan *"diastolik"*. Pemisahan tersebut dilakukan dengan asumsi bahwa nilai tekanan darah awalnya disimpan dalam format string yang dipisahkan oleh suatu karakter misalnya spasi. Kemudian, tipe data dari dua kolom baru ini diubah menjadi integer *(int32)* agar lebih sesuai untuk perhitungan numerik. Dalam kolom *"Diet"* melakukan pencetakan nilai unik yang berguna untuk melihat kategori - kategori apa saja yang ada dalam kolom tersebut, hal ini terdapat tiga kategori yaitu *"Average", "Unhealthy",* dan *"Healthy"*.

	Active Hours
0	17.384999
1	19.036541
2	14.536574
3	16.351019
4	22.485179
...	...
8758	13.193627
8759	20.166962
8760	21.624786
8761	23.970896
8762	14.994766

8763 rows × 1 columns

dtype: float64

Gambar 7. Pembuatan Kolom Baru *Active Hours*

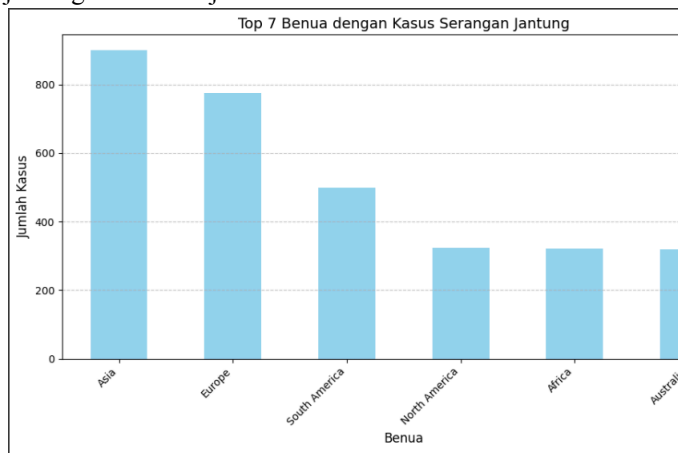
Persiapan untuk melakukan sebuah *Model Machine Learning* melakukan perhitungan total penggunaan memori oleh setiap kolom dalam *Data Frame*. Kemudian, total dalam penggunaan memori dari semua kolom dihitung dan ditampilkan dalam satuan byte dan kilobyte. Tujuan adalah untuk mendapatkan gambaran awal tentang seberapa besar memori yang digunakan oleh *Data Frame* tersebut. Kemudian melakukan iterasi pada setiap kolom dalam *Data Frame* yang diperlukan periksa. Jika tipe data dalam suatu kolom dapat diubah menjadi tipe data yang lebih hemat memori misalnya, dari *float64* ke *float16,* maka tipe data kolom tersebut akan diubah. Setelah semua perubahan tipe data yang dilakukan, penggunaan memori dihitung ulang untuk melihat seberapa besar penghematan yang telah dicapai. Penelitian ini diperlukan untuk mengkonversi *int32* menjadi *int16, float64* menjadi *float16*, dan *object* menjadi *category data*.

```
Memory before chaning datatype :  2558924
The memory after chsnging datatyos :  2506346
```

Gambar 8. Konversi Tipe Data

3.6 Data Manipulation and Visualization

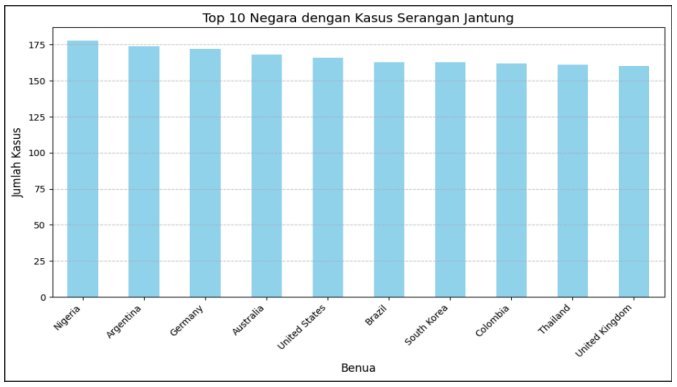
Penelitian ini diperlukan sebuah gambaran dalam dampak serangan jantung dan sudah beberapa terjadi karena adanya kasus masa lalu yang tercatat dalam tempat tempat di dalam dunia berbeda [6]. Berikut merupakan beberapa distribusi yang merupakan dari tipe Data Diagram Batang yang diharapkan untuk dapat memahami kasus serangan jantung terbesar terjadi.



Gambar 9. Data Visualisasi Kasus Serangan Jantung di Benua

Grafik di atas menyajikan data jumlah kasus serangan jantung di tujuh benua teratas. Terlihat bahwa Asia menempati posisi pertama dengan jumlah kasus serangan jantung tertinggi di antara benua lainnya. Posisi kedua ditempati oleh Eropa, diikuti oleh Amerika Selatan. Amerika Utara berada di posisi keempat, diikuti oleh Afrika dan Australia.

Analisis dari grafik tersebut, dapat disimpulkan bahwa terdapat perbedaan jumlah kasus serangan jantung yang cukup signifikan antar benua. Beberapa faktor seperti gaya hidup, pola makan, kondisi lingkungan dan akses terhadap pelayanan kesehatan kemungkinan besar berkontribusi terhadap perbedaan ini. Penting untuk melakukan penelitian lebih lanjut untuk mengidentifikasi faktor-faktor spesifik yang menyebabkan perbedaan ini dan mengembangkan strategi pencegahan yang efektif.

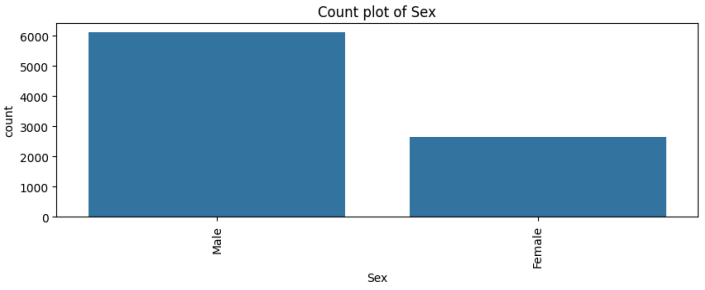


Gambar 10. Data Visualisasi Kasus Serangan Jantung dalam 10 Negara

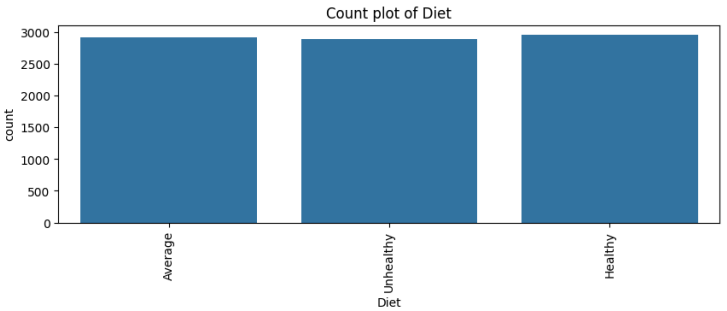
Visualisasi tersebut menampilkan 10 negara dengan jumlah kasus serangan jantung tertinggi. Berdasarkan visualisasi ini kita bisa dapat melihat bahwa Nigeria memiliki jumlah kasus serangan jantung tertinggi di antara 10 negara yang disajikan. Diikuti oleh Argentina dan Jerman yang memiliki jumlah kasus yang relatif dekat. Negara-negara lain seperti Australia, Amerika Serikat, Brasil, Korea Selatan, Kolombia, Thailand, dan Inggris juga mencatat jumlah kasus serangan jantung yang signifikan.

Secara umum, grafik ini menunjukkan bahwa ada variasi yang cukup besar dalam jumlah kasus serangan jantung di antara 10 negara. Meskipun Nigeria memiliki jumlah kasus tertinggi, tidak ada perbedaan yang signifikan di antara negara-negara lain. Hal ini menunjukkan bahwa serangan jantung merupakan masalah kesehatan global yang perlu mendapat perhatian serius di banyak negara. Perbedaan jumlah kasus dapat dipengaruhi oleh faktor-faktor seperti gaya hidup, pola makan, kondisi lingkungan, akses terhadap perawatan kesehatan, dan genetika.

Distribusi jenis kelamin dalam dataset menunjukkan ketidakseimbangan gender yang signifikan, dengan lebih banyak individu laki-laki daripada perempuan. Hal ini terlihat dari grafik batang, yang menunjukkan bahwa kategori *"Male"* memiliki tinggi batang yang lebih tinggi daripada kategori *"Female"*. Ketidakseimbangan ini mengindikasikan bahwa mayoritas data dalam dataset berasal dari individu laki-laki. Kondisi ini perlu diperhatikan, terutama jika dataset digunakan untuk analisis lebih lanjut, karena ketidakseimbangan gender dapat mempengaruhi generalisasi hasil analisis. Oleh karena itu, tindakan korektif, seperti pembobotan atau penyesuaian statistik, diperlukan untuk mengurangi potensi bias dan memastikan representasi yang lebih seimbang dalam analisis.



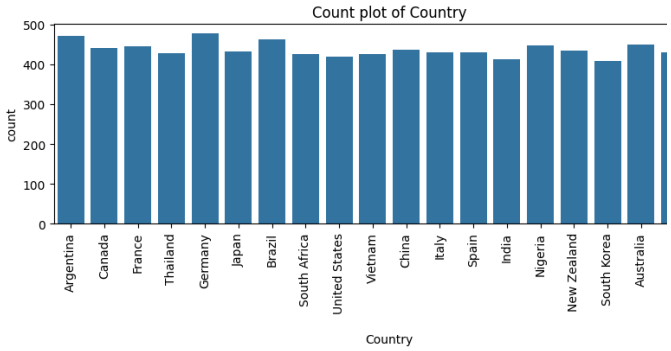
Gambar 11. Visualisasi Bar Chart Jenis Kelamin



Gambar 12. Visualisasi *Bar Chart Diet*

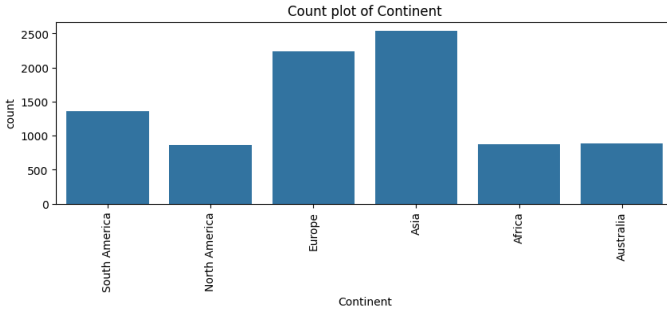
Diagram batang di atas menunjukkan distribusi pola makan dari sebuah set data. Dari grafik tersebut, dapat dilihat bahwa ada tiga kategori pola makan, yaitu *"Average"*, *"Unhealthy"*, dan *"Healthy"*. Jumlah individu dalam setiap kategori hampir sama, menunjukkan distribusi yang cukup seimbang. Hal ini mengindikasikan bahwa dalam dataset tersebut, terdapat proporsi yang seimbang antara individu yang memiliki pola makan rata-rata, tidak sehat, dan sehat. Namun, perlu dicatat bahwa kesimpulan ini

hanya didasarkan pada visualisasi data ini saja. Analisis lebih lanjut diperlukan untuk menarik kesimpulan yang lebih kuat dan memahami faktor-faktor yang mempengaruhi pola makan individu dalam dataset.



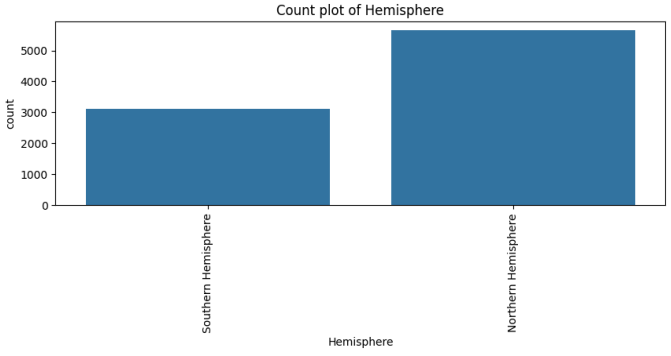
Gambar 13. Visualisasi *Bar Chart Country*

Pada grafik batang di atas menunjukkan distribusi frekuensi negara dalam kumpulan data. Ada sekitar 20 negara yang terwakili dalam data ini. Menariknya, distribusi negara dalam dataset ini cenderung merata. Tidak ada satu negara pun yang secara signifikan mendominasi jumlah data dibandingkan dengan negara lainnya. Hal ini mengindikasikan bahwa data yang dikumpulkan berasal dari berbagai negara dengan proporsi yang hampir sama.



Gambar 14. Visualisasi *Bar Chart Continent*

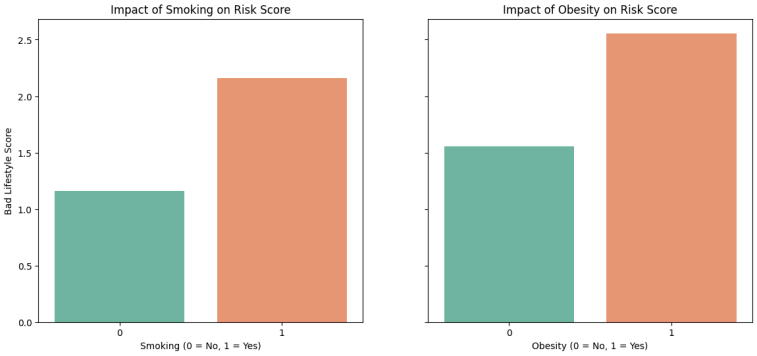
Gambar diagram batang di atas menunjukkan distribusi frekuensi data berdasarkan benua. Terlihat bahwa benua Asia memiliki jumlah data paling banyak dibandingkan dengan benua lainnya. Jumlah data di benua Eropa juga tinggi, diikuti oleh Amerika Selatan dan Afrika yang memiliki jumlah data yang relatif sama. Amerika Utara dan Australia memiliki jumlah data yang paling sedikit di antara benua lainnya. Dapat disimpulkan bahwa sebagian besar data berasal dari Asia dan Eropa, sedangkan data dari wilayah lain seperti Amerika Utara, Australia, Afrika, dan Amerika Selatan memiliki proporsi yang lebih kecil.



Gambar 15. Visualisasi *Bar Chart Hemisphere*

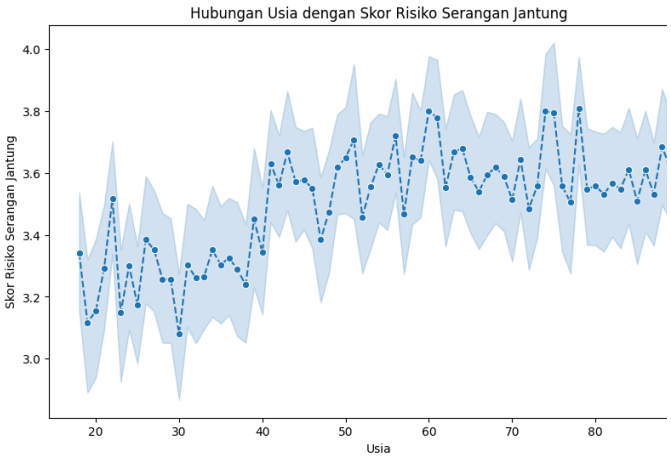
Pada grafik batang di atas menunjukkan distribusi data berdasarkan belahan bumi. Terlihat bahwa jumlah data dari belahan bumi utara jauh lebih tinggi daripada belahan bumi selatan. Hal ini mengindikasikan bahwa sebagian besar data yang dianalisis berasal dari negara-negara yang berada di belahan bumi utara. Ketidakseimbangan jumlah data dari kedua belahan bumi ini perlu diperhatikan ketika melakukan analisis lebih lanjut, karena dapat mempengaruhi hasil dan kesimpulan yang diperoleh. Kemungkinan besar, ketidakseimbangan ini dapat mempengaruhi generalisasi hasil, terutama jika seseorang ingin menyimpulkan sesuatu tentang populasi global secara keseluruhan. Untuk mendapatkan pemahaman yang lebih komprehensif, analisis lebih lanjut harus dilakukan dengan mempertimbangkan faktor-faktor lain yang relevan, seperti populasi di masing-masing belahan bumi, serta tujuan penelitian.

Penjelasan mengenai grafik visualisasi *Bar Chart* dalam menunjukkan hubungan tingkat resiko dalam gaya hidup secara individu yang menunjukkan kebiasaan dalam merokok dan obesitas terukur dalam visualisasi ini secara singkatnya Grafik sebelah kiri menunjukkan bahwa individu yang merokok (nilai 1) memiliki skor risiko yang lebih tinggi dibandingkan dengan yang tidak merokok (nilai 0). Hal ini mengindikasikan bahwa merokok berkontribusi terhadap peningkatan risiko gaya hidup yang tidak sehat. Grafik sebelah kanan menunjukkan hasil yang serupa, dimana individu yang mengalami obesitas (nilai 1) memiliki skor risiko yang lebih tinggi secara signifikan dibandingkan dengan mereka yang tidak mengalami obesitas (nilai 0). Hal ini mengindikasikan bahwa obesitas juga merupakan faktor risiko yang signifikan untuk gaya hidup tidak sehat.



Gambar 16. Visualisasi Dampak Gaya Hidup Merokok dan Obesitas dalam Resiko Serangan Jantung

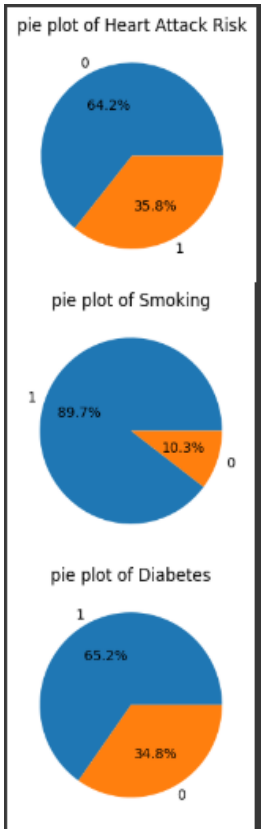
Dampak usia dalam resiko serangan jantung juga berpengaruh karena umur merupakan faktor yang termasuk dalam sebuah faktor yang bisa meningkatkan dalam serangan jantung dengan visualisasi ini bisa memberikan sebuah representasi dalam sebuah skor resiko yang dimiliki dalam setiap usia.



Gambar 17. Visualisasi Line Plot Diagram Hubungan Usia dengan Skor Resiko Serangan Jantung

Garis putus-putus biru menunjukkan skor risiko rata-rata untuk setiap kelompok usia, sedangkan area biru muda di sekitarnya menunjukkan rentang atau interval kepercayaan skor risiko. Berdasarkan grafik ini, dapat dilihat bahwa secara umum, skor risiko serangan jantung cenderung meningkat seiring bertambahnya usia. Namun, terdapat fluktuasi yang cukup besar di dalam setiap kelompok usia, yang menunjukkan bahwa usia bukanlah satu-satunya faktor yang mempengaruhi risiko serangan jantung.

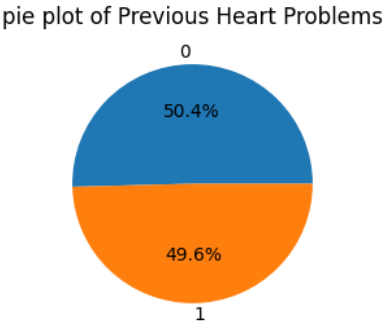
Persebaran Distribusi data mempunyai macam - macam dalam individu mempunyai dampak yang bisa meningkatkan sebuah resiko dalam serangan penyakit jantung inilah beberapa diagram Pie chart untuk jumlah dalam jenis variabel kolom pada dataset ini.



Gambar 18. Visualisasi Diagram Pie Chart *Smoking*, *Diabetes*, dan *Heart Attack Risk*

Diagram Pie Chart dalam dataset ini menggambarkan proporsi atau persentase dari dua kategori pada tiga variabel utama:heart attack risk, smoking habits, and diabetic conditions. Berdasarkan analisis, 64,2% individu memiliki risiko serangan jantung yang rendah, sementara 35,8% memiliki risiko yang lebih tinggi, yang mengindikasikan bahwa proporsi individu dengan risiko tinggi cukup signifikan. Pada variabel kebiasaan merokok, 89,7% individu adalah bukan perokok, sementara hanya 10,3% yang merupakan perokok, yang mengindikasikan bahwa mayoritas sampel adalah individu yang tidak merokok. Untuk kondisi diabetes, 65,2% individu dalam dataset menderita diabetes, sementara 34,8% tidak menderita diabetes.

Penjelasan secara singkat untuk diagram lingkaran Pie Chart “Previous Heart Problems” menunjukkan distribusi individu berdasarkan riwayat masalah jantung sebelumnya, dengan hasil yang hampir sama di antara kedua kategori. Sebanyak 50,4% individu memiliki riwayat masalah jantung, sementara 49,6% tidak memiliki riwayat tersebut. Proporsi yang hampir sama ini menunjukkan bahwa dalam sampel data yang dianalisis, terdapat representasi yang seimbang antara individu yang memiliki dan yang tidak memiliki riwayat masalah jantung. Temuan ini penting karena memberikan gambaran yang lebih merata mengenai prevalensi riwayat masalah jantung dalam populasi, yang dapat menjadi dasar untuk analisis lebih lanjut mengenai faktor risiko.



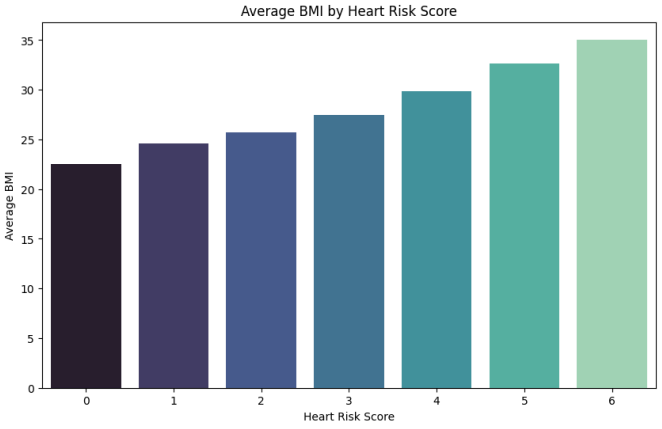
Gambar 19. Visualisasi Diagram Pie Chart *Previous Heart Problems*

Gambar 20. Visualisasi Histogram dalam Distribusi Skor Gaya Hidup Buruk dengan Riwayat Serangan Jantung

Berdasarkan grafik di atas, terlihat distribusi skor gaya hidup yang buruk pada dua kelompok individu, yaitu kelompok yang memiliki riwayat serangan jantung dan

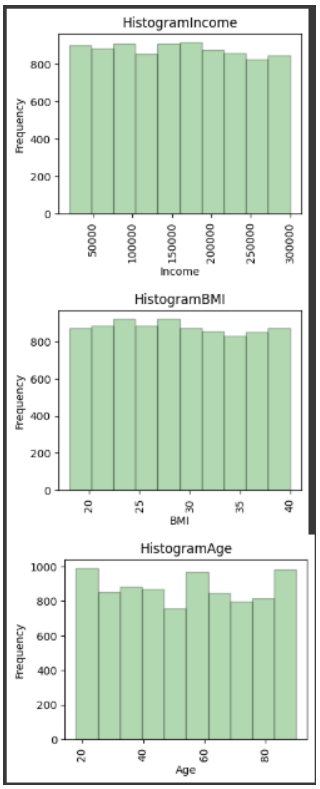
kelompok yang tidak memiliki riwayat serangan jantung. Dapat dilihat bahwa secara umum, baik kelompok dengan dan tanpa riwayat serangan jantung memiliki distribusi skor gaya hidup buruk yang serupa, yang cenderung berpusat di sekitar skor 2,0. Hal ini mengindikasikan bahwa sebagian besar individu pada kedua kelompok memiliki skor gaya hidup yang buruk di sekitar nilai ini. Namun, ada sedikit perbedaan antara kedua kelompok, di mana kelompok dengan riwayat serangan jantung cenderung memiliki sedikit lebih banyak individu dengan skor gaya hidup buruk yang tinggi (mendekati 4,0) dibandingkan dengan kelompok tanpa riwayat serangan jantung. Kondisi ini menunjukkan bahwa meskipun distribusi umum dari kedua kelompok serupa, individu dengan riwayat serangan jantung cenderung memiliki gaya hidup yang sedikit lebih tidak sehat dibandingkan dengan mereka yang tidak memiliki riwayat serangan jantung.

Penjelasan mengenai gambar grafik batang di bawah menunjukkan hubungan antara *Heart Risk Score* dan rata-rata indeks massa tubuh (BMI). Semakin tinggi skor risiko jantung, semakin tinggi pula rata-rata BMI individu. Hal ini menunjukkan tren peningkatan BMI seiring dengan meningkatnya risiko penyakit jantung. Dengan kata lain, individu dengan risiko jantung yang lebih tinggi cenderung memiliki berat badan yang lebih tinggi dibandingkan dengan mereka yang memiliki risiko jantung yang lebih rendah.



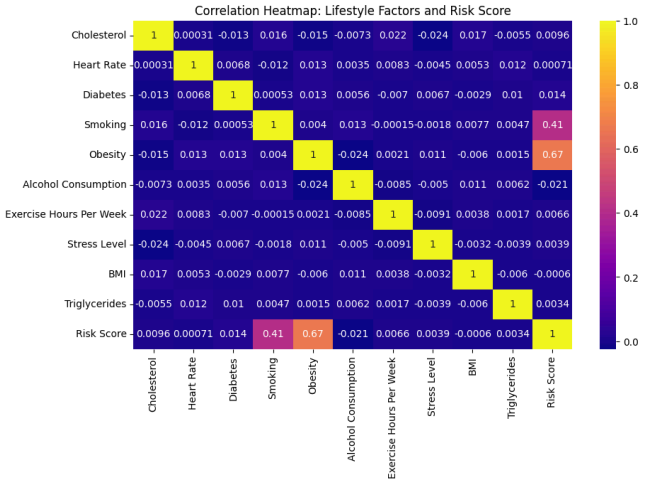
Gambar 21. Visualisasi dalam Distribusi Bar Chart Average BMI by Heart Risk Score

Histogram yang menggambarkan distribusi frekuensi dari tiga variabel yaitu *income*, indeks massa tubuh (BMI), dan *age*, memberikan wawasan penting tentang karakteristik populasi dalam dataset. Distribusi pendapatan menunjukkan bahwa sebagian besar individu memiliki *income* antara 50.000 dan 250.000, dengan frekuensi yang cenderung merata pada kisaran ini, tetapi menurun pada kelompok pendapatan tertinggi. Untuk BMI, sebagian besar individu memiliki BMI antara 20 hingga 35, yang mengindikasikan bahwa mayoritas penduduk memiliki berat badan normal atau sedikit kelebihan berat badan, dengan frekuensi yang lebih rendah pada kelompok BMI yang sangat rendah atau sangat tinggi. Sementara itu, distribusi *age* menunjukkan bahwa data mencakup rentang usia yang luas antara 20 hingga 80 tahun, dengan distribusi yang relatif merata di sebagian besar rentang usia, meskipun terdapat penurunan frekuensi pada kelompok usia sangat muda dan sangat tua. Temuan ini memberikan gambaran yang beragam tentang aspek sosial dan kesehatan dalam set data yang dianalisis.



Gambar 22. Visualisasi Histogram Hitungan Dataset dalam Variabel *Income,* BMI, dan *Age*

3.7 Menganalisis Korelasi dalam Data



Gambar 23. Visualisasi Heatmap dalam Hubungan Data Variabel Korelasi dalam Semua Faktor Gaya Hidup dan *Risk Score*

Pada Visualisasi Correlation Heatmap antara variabel gaya hidup dan skor risiko penyakit jantung memberikan gambaran tentang kekuatan dan arah hubungan antar variabel. Warna yang lebih terang, seperti kuning terang, menunjukkan korelasi positif yang kuat, sedangkan warna yang lebih gelap, seperti ungu tua, menunjukkan korelasi negatif yang kuat. Berdasarkan Correlation Heatmap ini, variabel-variabel seperti kolesterol, tekanan darah, diabetes, merokok, obesitas, dan trigliserida menunjukkan korelasi positif yang signifikan dengan skor risiko penyakit jantung, yang berarti peningkatan nilai variabel-variabel ini dikaitkan

dengan peningkatan risiko penyakit jantung. Sebaliknya, konsumsi alkohol dan jumlah jam olahraga per minggu memiliki korelasi negatif yang lemah dengan skor risiko penyakit jantung, yang menunjukkan bahwa aktivitas olahraga dan konsumsi alkohol dalam batas-batas tertentu dapat sedikit menurunkan risiko. Tingkat stres menunjukkan korelasi positif yang lemah, sementara BMI (indeks massa tubuh) memiliki korelasi positif yang sedang dengan skor risiko, menandakan obesitas sebagai faktor risiko yang cukup signifikan. Secara keseluruhan, analisis ini menegaskan hubungan antara berbagai faktor gaya hidup dan risiko penyakit jantung, meskipun penting untuk dicatat bahwa korelasi tidak selalu menunjukkan hubungan sebab akibat. Faktor-faktor lain yang tidak dianalisis dalam peta panas ini mungkin juga berperan dalam menentukan risiko penyakit jantung.

IV. Hasil dan Pembahasan

Tahapan ini merupakan pembahasan mengenai *Model Machine Learning* menggunakan *Logistic Regression* (LR) dan *Gaussian Naive Bayes* (GNB) untuk Memprediksi dalam kasus yang terkenanya serangan jantung dalam jumlah 8.763 kasus. Berikut adalah Hasil dari *Model Machine Learning* tersebut.

4.1 Logistic Regression

Logistic Regression Report:

	precision	recall	f1-score	support
0	0.63	0.99	0.77	1123
1	0.98	0.41	0.58	1127
accuracy			0.70	2250
macro avg	0.80	0.70	0.67	2250
weighted avg	0.80	0.70	0.67	2250

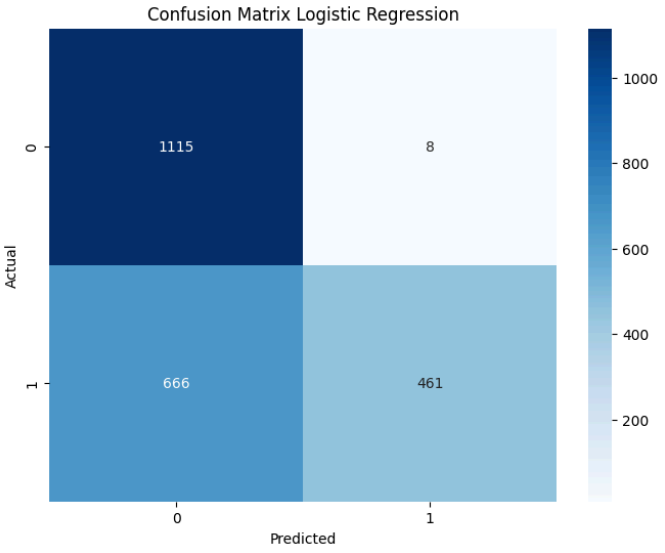
Accuracy: 0.7004444444444444

Gambar 24. Akurasi Model *Logistic Regression*

Model *Logistic Regression* ini memiliki akurasi sebesar 70% (0.7004444444444444). Akurasi menunjukkan persentase prediksi yang benar dibandingkan dengan total data yang diuji. Meskipun akurasi memberikan gambaran umum tentang kinerja model, metrik ini saja tidak cukup untuk menilai kualitas model, terutama jika data tidak seimbang di antara kelas-kelasnya. Kemudian Presisi mengukur seberapa baik model dalam membuat prediksi positif yang benar, yaitu rasio antara True Positives dan total prediksi positif (True Positives + False Positives). Presisi untuk kelas 0 adalah 0,63, yang berarti bahwa hanya 63% dari prediksi kelas 0 yang benar-benar benar. Sementara itu, presisi untuk kelas 1 adalah 0.98, yang menunjukkan bahwa hampir semua prediksi untuk kelas 1 akurat. Hal ini mengindikasikan bahwa model memiliki tingkat kesalahan yang rendah dalam memprediksi kelas 1.

Selanjutnya *Recall* mengukur kemampuan model untuk mendeteksi semua data aktual yang positif, yang merupakan rasio antara True Positives dan total data aktual yang positif (True Positives + False Negatives). *Recall* untuk kelas 0 sangat tinggi yaitu 0.99, yang berarti bahwa hampir semua data kelas 0 diidentifikasi dengan benar. Sebaliknya, *recall* untuk kelas 1 cukup rendah yaitu 0.41, yang mengindikasikan bahwa model kurang mampu mendeteksi sebagian besar data aktual dari kelas 1. *F1-score* yang merupakan rata-rata harmonik antara presisi dan recall, memberikan keseimbangan antara kedua metrik tersebut. *F1-score* untuk kelas 0 adalah 0.77, yang mencerminkan kinerja yang cukup baik pada kelas ini. Namun, untuk kelas 1, *F1-score* hanya 0.58, yang menunjukkan bahwa model tidak cukup optimal dalam menangani kelas data ini, terutama karena rendahnya recall.

Support mengacu pada jumlah data aktual untuk setiap kelas dalam dataset. Dalam laporan ini, terdapat 1123 data untuk kelas 0 dan 1127 data untuk kelas 1. Jumlah data yang seimbang antara kedua kelas ini memungkinkan evaluasi model yang lebih adil, meskipun kinerja model pada kelas 1 masih jauh dari memadai. *Macro Avg* memberikan nilai rata-rata presisi, recall, dan F1-score tanpa mempertimbangkan proporsi setiap kelas dalam dataset. Nilai *macro avg* menunjukkan precision 0.80, recall 0.70, dan F1-score 0.67. Sementara itu, weighted avg memberikan rata-rata yang memperhitungkan jumlah data setiap kelas. Nilai weighted avg sama dengan macro avg pada laporan ini karena jumlah data antara kelas 0 dan 1 hampir sama.



Gambar 25. *Confusion Matrix Logistic Regression*

Confusion matrix di atas menunjukkan performa model Regresi Logistik dalam memprediksi data uji. True Positives (TP) sebesar 461 dan True Negatives (TN) sebesar 1115 menunjukkan bahwa model cukup baik dalam mengidentifikasi data kelas 1 dan kelas 0 dengan benar, terutama untuk kelas 0. False Positives (FP) hanya 8, yang berarti hanya sedikit data kelas 0 yang salah diprediksi sebagai kelas 1, sedangkan False Negatives (FN) sebanyak 666 menunjukkan bahwa model sering salah memprediksi data kelas 1 sebagai kelas 0, yang berakibat pada rendahnya recall untuk kelas 1. Dengan total data sebanyak 2250, akurasi model adalah 70% (0.7004), yang dihitung dari rasio prediksi yang benar (TP+TN) terhadap total data. Namun, nilai FN yang tinggi mengindikasikan bahwa model kesulitan untuk mendeteksi data kelas 1, meskipun performa pada kelas 0 sangat baik. Untuk meningkatkan performa, terutama dalam mendeteksi kelas 1, strategi seperti menyeimbangkan dataset, menggunakan metrik alternatif

seperti F1-score, atau mencoba algoritma lain yang lebih sesuai dapat diterapkan.

4.2 Gaussian Naive Bayes

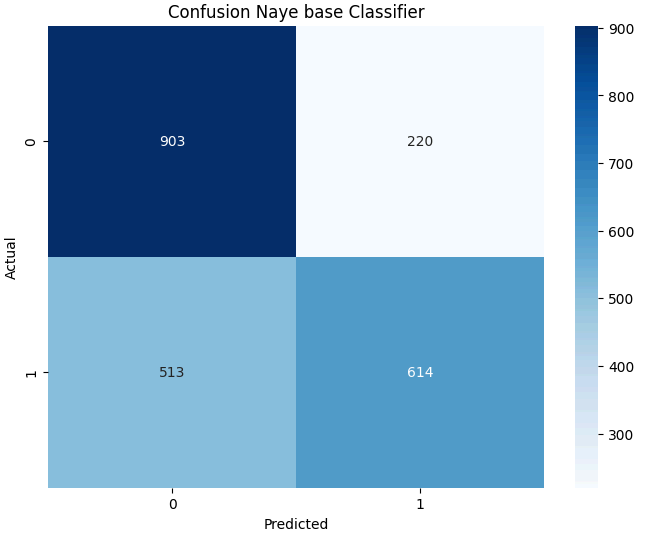
```
Naye base Classifier:
              precision    recall  f1-score   support

           0       0.64      0.80      0.71      1123
           1       0.74      0.54      0.63      1127

    accuracy                           0.67      2250
   macro avg       0.69      0.67      0.67      2250
weighted avg       0.69      0.67      0.67      2250

Accuracy: 0.6742222222222222
```

Gambar 26. Akurasi Model *Gaussian Naive Bayes*

Hasil Akurasi pada gambar di atas mengevaluasi kinerja model Naïve Bayes Classifier, dengan akurasi keseluruhan 67,42% (0.6742222222222222). Akurasi ini dihitung sebagai rasio antara jumlah prediksi yang benar (True Positif dan True Negatif) dan jumlah total data, yaitu 2250. Meskipun akurasi memberikan gambaran umum, menganalisis metrik lain seperti presisi, recall, dan F1-score sangat penting untuk mengevaluasi kinerja model secara lebih rinci, terutama ketika mempertimbangkan kinerja setiap kelas.

Presisi untuk kelas 0 adalah 0,64, yang menunjukkan bahwa 64% dari prediksi untuk kelas 0 benar. Recall untuk kelas 0 lebih tinggi yaitu 0.80, yang berarti bahwa model mampu mendeteksi 80% dari data kelas 0 yang sebenarnya dengan benar. Kombinasi presisi dan recall menghasilkan skor F1 sebesar 0,71 untuk kelas 0, yang mencerminkan keseimbangan antara kemampuan model untuk membuat prediksi yang benar dan mendeteksi data aktual kelas ini. Sebaliknya, untuk kelas 1, presisi lebih tinggi yaitu 0,74, yang menunjukkan bahwa sebagian besar prediksi kelas 1 benar. Namun, recall untuk kelas 1 hanya 0,54, yang berarti bahwa hanya 54% dari data kelas 1 yang sebenarnya berhasil diidentifikasi oleh model. Recall yang rendah ini mengindikasikan bahwa model sering salah memprediksi data kelas 1 sebagai kelas 0. Nilai F1-score untuk kelas 1 adalah 0.63, lebih rendah dari kelas 0.

Berdasarkan keseluruhan, rata-rata makro presisi, recall, dan F1-score masing-masing adalah 0,69, 0,67, dan 0,67. Rata-rata makro memberikan rata-rata sederhana dari metrik untuk kedua kelas tanpa mempertimbangkan distribusi data. Rata-rata tertimbang, yang memperhitungkan proporsi data di setiap kelas, memiliki nilai yang sama yaitu 0,69 untuk presisi dan 0,67 untuk recall dan F1-score. Nilai rata-rata tertimbang, yang mirip dengan rata-rata makro, disebabkan oleh distribusi data yang hampir seimbang antara kelas 0 dan kelas 1, dengan masing-masing 1123 dan 1127 data. Pada secara umum, model Naïve Bayes memiliki performa yang lebih baik dalam mendeteksi data dari kelas 0 dibandingkan kelas 1. Hal ini terlihat dari recall dan F1-score yang lebih tinggi untuk kelas 0. Rendahnya recall untuk kelas 1 mengindikasikan bahwa model sering gagal untuk mengidentifikasi data aktual dari kelas ini, yang merupakan kelemahan yang signifikan. Untuk meningkatkan kinerja model, langkah-langkah seperti penyesuaian parameter, penghapusan fitur yang kurang relevan, atau metode penyeimbangan data seperti oversampling untuk kelas 1 dapat diterapkan. Hal ini dapat membantu model mendeteksi dan memprediksi data kelas 1 dengan lebih baik tanpa mengorbankan kinerja pada kelas 0.



Gambar 27. *Confusion Matrix Gaussian Naye Base*

Gambar 27 di atas ini adalah confusion matrix dari model klasifikasi menggunakan algoritma Naïve Bayes. Hasilnya menunjukkan bahwa model tersebut berhasil memprediksi 614 sampel dari kelas “1” dengan benar (True Positive) dan 903 sampel dari kelas “0” dengan benar (True Negative). Namun, model tersebut juga menghasilkan 220 kesalahan dengan memprediksi kelas “0” sebagai “1” (False Positive) dan 513 kesalahan dengan memprediksi kelas “1” sebagai “0” (False Negative). Hasil ini menunjukkan bahwa model lebih baik dalam mengenali kelas “0” daripada kelas “1”, seperti yang terlihat dari tingginya jumlah prediksi True Negative.

Walaupun secara keseluruhan model memiliki performa yang cukup baik, tingginya jumlah kesalahan False Positive dan False Negative mengindikasikan bahwa model memiliki keterbatasan dalam membedakan kelas “0” dan “1”. Hal ini dapat disebabkan oleh ketidakseimbangan data atau kompleksitas pola data yang tidak dapat sepenuhnya dipahami oleh model. Oleh karena itu, untuk meningkatkan kinerja model, langkah-langkah seperti penyesuaian parameter, menggunakan data yang lebih seimbang, atau mencoba algoritma klasifikasi lain dapat dipertimbangkan. Selain itu, analisis tambahan seperti presisi, recall, dan F1-score dapat memberikan wawasan yang lebih dalam untuk mengevaluasi kemampuan model secara menyeluruh.

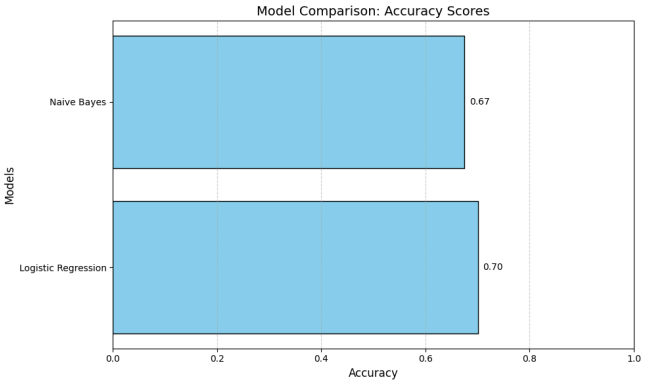
4.3 Model Evaluasi dan Perbandingan

Gambar dibawah ini tersebut menunjukkan perbandingan metrik evaluasi antara dua algoritma klasifikasi, yaitu *Gaussian Naïve Bayes* (GaussianNB) dan Regresi Logistik. Regresi Logistik menunjukkan akurasi yang lebih tinggi (70,04%) dibandingkan GaussianNB (67,42%), yang mengindikasikan kinerjanya yang sedikit lebih baik secara

keseluruhan. Dari segi presisi, *Logistic Regression* juga unggul dengan nilai 98.29%, jauh di atas GaussianNB yang hanya 73.62%, mengindikasikan bahwa *Logistic Regression* lebih dapat diandalkan dalam meminimalisir jumlah false positive. Sebaliknya, GaussianNB memiliki *recall* yang lebih tinggi (54,48%) daripada Regresi Logistik (40,91%), yang menunjukkan kemampuannya yang lebih baik dalam mendeteksi sampel positif sebanyak mungkin, meskipun dengan risiko positif palsu yang lebih besar. GaussianNB juga memiliki F1 Score yang lebih tinggi (62,62%) daripada Regresi Logistik (57,77%), yang mencerminkan keseimbangan yang lebih baik antara presisi dan recall. Secara keseluruhan, Regresi Logistik lebih cocok untuk kasus-kasus yang memprioritaskan prediksi positif yang benar dan akurasi yang tinggi, sedangkan GaussianNB lebih cocok untuk situasi yang membutuhkan deteksi positif sebanyak mungkin meskipun dengan *trade-off* pada kesalahan prediksi. Pilihan algoritma tergantung pada tujuan spesifik dari analisis yang dilakukan.

	GaussianNB	LogisticRegression
Accuracy	67.420000	70.040000
F1 Score	62.620000	57.770000
Precision	73.620000	98.290000
recall	54.480000	40.910000

Gambar 28. Gambar Perbandingan Model Akurasi



Gambar 29. Visualisasi Bar Chart Perbandingan Akurasi

Pada Gambar diatas menunjukkan menunjukkan perbandingan nilai akurasi antara dua model, Naïve Bayes dan Regresi Logistik. Regresi Logistik memiliki akurasi sebesar 0.70 atau 70%, sedikit lebih tinggi dibandingkan dengan Naïve Bayes dengan akurasi 0.67 atau 67%. Perbedaan ini menunjukkan bahwa Regresi Logistik lebih unggul dalam memprediksi dengan benar secara keseluruhan dibandingkan dengan Naïve Bayes. Namun, perbedaan akurasi ini tidak terlalu signifikan, sehingga pemilihan model sebaiknya tidak hanya bergantung pada akurasi saja, tetapi juga pada metrik evaluasi lainnya, seperti precision, recall, dan F1 Score, yang dapat memberikan gambaran yang lebih lengkap mengenai performa kedua model tersebut.

V. KESIMPULAN

Machine learning adalah cabang dari kecerdasan buatan yang berfokus pada pengembangan algoritma dan model yang mampu mempelajari pola dalam data untuk membuat prediksi atau keputusan tanpa harus diprogram secara eksplisit. Dalam konteks penelitian ini, dua algoritma, yaitu Naïve Bayes dan Regresi Logistik, digunakan untuk memprediksi risiko serangan jantung berdasarkan variabel gaya hidup dan kesehatan. Hasil evaluasi menunjukkan bahwa Regresi Logistik memiliki akurasi yang sedikit lebih tinggi (70%) dibandingkan dengan Naïve Bayes (67%). Perbedaan ini mengindikasikan bahwa Regresi Logistik mampu memberikan prediksi yang lebih baik secara keseluruhan, meskipun perbedaan akurasi tersebut tidak terlalu signifikan.

Hasil evaluasi ini menegaskan bahwa pemilihan model dalam machine learning sebaiknya diperlukan mendapatkan akurasi yang sesuai untuk memastikan akurasi tersebut tercapai kriteria. Hal ini penting untuk memastikan bahwa model yang dipilih tidak hanya mampu memprediksi dengan benar secara keseluruhan, tetapi juga efektif dalam menangani berbagai jenis kesalahan, terutama dalam kasus klasifikasi risiko penyakit yang dapat berdampak pada kesehatan manusia.

Pada penelitian ini, analisis faktor risiko serangan jantung menunjukkan bahwa gaya hidup dan kesehatan memiliki peran yang signifikan. Variabel-variabel seperti kolesterol, tekanan darah, diabetes, merokok, obesitas, dan trigliserida memiliki korelasi positif dengan risiko serangan jantung. Hal ini menunjukkan bahwa individu dengan nilai yang lebih tinggi pada variabel-variabel ini cenderung memiliki risiko serangan jantung yang lebih besar. Sebaliknya, faktor-faktor seperti olahraga dan konsumsi alkohol dalam jumlah tertentu memiliki korelasi negatif, meskipun lemah, yang mengindikasikan adanya peran protektif terhadap risiko serangan jantung.

Kemudian dalam hasil menganalisis data menunjukkan bahwa stres dan BMI (indeks massa tubuh) memiliki pengaruh yang moderat terhadap risiko penyakit. Tingkat stres yang tinggi dan BMI yang tinggi (obesitas) dapat meningkatkan kemungkinan serangan jantung. Oleh karena itu, manajemen stres dan pemeliharaan berat badan yang sehat merupakan strategi penting dalam pencegahan penyakit kardiovaskular. Kesadaran akan faktor-faktor risiko ini penting untuk mendukung intervensi berbasis gaya hidup yang lebih efektif.

Secara keseluruhan, penelitian ini menyoroti pentingnya penggunaan machine learning dalam analisis data kesehatan untuk mendukung pengambilan keputusan yang lebih baik. Dengan mempertimbangkan hasil akurasi dan evaluasi model, serta hubungan antara gaya hidup dan risiko penyakit, hasil ini dapat digunakan untuk mengembangkan program pencegahan yang lebih tepat sasaran. Penggunaan teknologi ini memberikan peluang besar untuk meningkatkan pemahaman kita tentang faktor-faktor yang mempengaruhi kesehatan, sehingga membantu mengurangi beban penyakit kardiovaskular di masyarakat.

REFERENCES

[1] Bagus Prayogi, M., Irawan, I., Fajar, Y. I., & Kunci, K. (2024). Perbandingan Algoritma ID3, Naive Bayes, SVM Berbasis PSO Untuk Prediksi Serangan Jantung. The 3rd MDP Student Conference 2024, Vol 3 No 1 https://doi.org/10.35957/mdp-sc.v3i1.6979.

[2] D. Saputra, W. Irmayani, D. Purwaningtyas, and J. Sidauruk, "Comparative Analysis of Classification Algorithms C4.5, Naïve Bayes and Support Vector Machine Based on Particle Swarm Optimization (PSO) for Heart Disease Prediction," Int. J. Adv. Data Inf. Syst., vol. 2, no. 2, pp. 84-95, 2021, doi: 10.25008/ijadis.v2i2.1221.

[3] D. Maulana, "Implementasi Algoritma Naïve Bayes Untuk Klasifikasi Penderita Penyakit Jantung Di Indonesia Menggunakan Rapid Miner," Vol. 10, pp. 191–197, 2019.

[4] Nugraha, W., Sabaruddin, R., & Murni, S. (2024). Teknik Penskalaan Menggunakan Robust Scaler untuk Mengatasi Data Outlier pada Model Prediksi Serangan Jantung (Vol. 23, Issue 2). doi 10.62411/tc.v23i2.10463

[5] I. Harifal dan L. D. Lestari, "Optimasi Naive Bayes dengan PSO untuk Prediksi Kebutuhan ICU Pasien Covid-19," Sistemasi, Vol. 11, No. 3, pp. 724-734, 2022.

[6] M. Ozcan and S. Peker, "Classification and regression tree algorithms for heart disease modeling and prediction of heart disease," Healthcare Analytics, vol. 3, no. December 2022, pp. 100130, 2023, doi: 10.1016/j.health.2022.100130.

[7] P. P. Ghosh et al, "Efficient cardiovascular disease prediction using machine learning with relief and lasso feature selection techniques," IEEE Access, vol. 9, pp. 19304-19326, 2021, doi:10.1109/ACCESS.2021.3053759.

[8] Santoso, M. et al, (2023) 'Klasifikasi Potensi Penyakit Jantung Menggunakan Algoritma C4. 5 Algorithm', Jurnal INSAN Manajemen Sistem Informasi Inovasi, 3(2), pp. 96-103.

[9] Khalisatifa Aida, Arum Hestiana Dela, & Jambak Muhammad Ihsan. (2024). RISK CLASSIFICATION OF HEART ATTACK DISEASE USING C4.5 ALGORITHM. Scientific Journal of Computers and Technology, Vol14 No 1, 57-64 https://doi.org/10.32699/device.v14i1.6869.

[10] F. Ali et al., "Smart health monitoring system for heart disease prediction based on ensemble deep learning and feature fusion, ensemble deep learning and feature fusion," Information Fusion, vol. 63, pp. 208-222, 2020, doi: 10.1016/j.inffus.2020.06.008.

[11] Bianto, M.A., Kusrini, K., & Sudarmawan, S. "Rancang Bangun Sistem Klasifikasi Penyakit Jantung Menggunakan Naïve Bayes." Jurnal Teknologi Informasi Kreatif, vol. 6, no. 1, pp. 75-83, 2020.

[12] Agus Oka Gunawan, I.M. dkk. (2023) 'Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Seri C4.5 dengan Rapidminer', Jurnal Teknologi dan Sistem Informasi Bisnis, 5(2), pp. 73-83. https://doi.org/10.47233/jteksis.v5i2.775.

[13] Pangaribuan, J. J., Tanjaya, H., & Kenichi, K. "Pendeteksian Penyakit Jantung Menggunakan Machine Learning dengan Algoritma Regresi Logistik." Jurnal Pengembangan Sistem Informasi (ISD), vol. 6, no. 2, pp. 1-10, 2021.

[14] Tasia, E. et al, (2023) Klasifikasi Penyakit Gagal Jantung Menggunakan Pembelajaran Terawasi. https://journal.irpi.or.id/index.php/sentimas.

[15] Li, W., Wang, C.-H., Cheng, G., & Song, Q. (2021). Optimal-statistical Collaboration Towards Generalized and Efficient Black-box Optimization. http://arxiv.org/abs/2106.09215

[16] Mahesh, B. (2020). Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR), 9(1), 381-386. https://doi.org/10.21275/art20203995

[17] Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, Vol 31, 685-695. https://doi.org/10.1007/s12525-021-00475-2/Published

[18] Schober, Patrick MD, PhD, MMedStat*; Vetter, Thomas R. MD, MPH†. Logistic Regression in Medical Research. Anesthesia & Analgesia 132(2): p 365-366, February 2021. DOI: 10.1213/ANE.0000000000005247

[19] As'ad, I. (2023). Advancing Health Diagnostics: A Study on Gaussian Naive Bayes Classification of Blood Samples. International Journal of Artificial Intelligence in Medical Problems, 1(2), 115-123. https://doi.org/10.56705/ijaimi.v1i2.120

[20] S. Naiem, A. E. Khedr, A. M. Idrees and M. I. Marie, "Improving the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in Detecting DDOS in Cloud Computing," in IEEE Access, vol. 11, pp. 124597-124608, 2023, doi: 10.1109/ACCESS.2023.3328951

[21] Anandhini M. N., Fanindia P., Marischa E., Chatarina F., Niskarto Z., Umaya R. P. N., & Romi F. R. (2024). Logistic Regression for Merchant Customer Churn Prediction: A Data-Driven Approach. Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI) (MLMI '24). Association for Computing Machinery, New York, NY, USA, 72-77. https://doi.org/10.1145/3696271.3696283

[22] Radchenko, O. V., Pavlov, V. A., Horodetska, O. K., & Korniienko, G. A. (2023). A multiclass classifier based on binary logistic regression obtained based on GMDH principles. Control and Computer Systems, 3(3), 24-32. https://doi.org/10.15407/csc.2023.03.024

[23] Bisong, E. (2019). Logistic Regression. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_20

[24] Das, A. (2023). Logistic Regression. In: Maggino, F. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Cham. https://doi.org/10.1007/978-3-031-17299-1_1689

[25] Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J., Kull, M., Lachiche, N., Quintana, M., & Flach, P. (2021). CRISP-DM Dua Puluh Tahun Kemudian: Dari Proses Penggalian Data ke Lintasan Ilmu Data. IEEE Transactions on Knowledge and Data Engineering, 33, 3048-3061. https://doi.org/10.1109/tkde.2019.2962680.

[26] Vanegas, C., Mejía, J., Agudelo, F., & Duran, D. (2023). Essence-based representation for the CRISP-DM methodology. Computación y Sistemas (CyS), 27. https://doi.org/10.13053/cys-27-3-3446.

[27] Krishnaswamy, V., Singh, N., Sharma, M., Verma, N., & Verma, A. (2022). Application of CRISP-DM methodology to manage human-wildlife conflict: an empirical case study in India. Journal of Environmental Planning and Management, 66, 2247 - 2273. https://doi.org/10.1080/09640568.2022.2070460.

[28] Schröer, C., Kruse, F., & Gómez, J. (2020). A Systematic Literature Review on the Application of the CRISP-DM Process Model. , 526-534. https://doi.org/10.1016/J.PROCS.2021.01.199.

[29] Wang, W. (2024). Data-driven analysis of carbon emissions from building construction under CRISP-DM framework. Applied Mathematics and Nonlinear Science, 9. https://doi.org/10.2478/amns-2024-1778.

[30] Sukhdeve, D.S.R., Sukhdeve, S.S. (2023). Google Colaboratory. Dalam: Google Cloud Platform for Data Science. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-9688-2_2