CENTIVE
VOL. 4, NO. 1, PP.1102-1113, DEC 2024
ISBN: 978-602-53004-0-0

Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media 2024

Implementasi Penggunaan Model Regresi Linier untuk Memprediksi Risiko Penyakit Jantung Berdasarkan Data Medis

Annisa Dwi Rahmawati[1] , Putri Eka Nur Alifah[2] , Dela Setiowati*[3]

Sistem Informasi, Rekayasa Industri, Universitas Telkom Purwokerto
Jl. DI Panjaitan No.128, Karangreja, Purwokerto Kidul, Kec. Purwokerto Sel., Kabupaten Banyumas, Jawa Tengah 53147, Indonesia

[1] annisadr@student.telkomuniversity.ac.id
[2] putriekanuralifah@student.telkomuniversity.ac.id
[3] delasetiowati@student.telkomuniversity.ac.id

Dikirim pada 17-11-2024, Direvisi pada 26-11-2024, Diterima pada 30-11-2024

Abstrak

Penyakit jantung menjadi tantangan yang serius yang mempengaruhi jutaan orang di seluruh dunia. Penyakit ini dapat menyerang organ vital yang memompa darah ke seluruh tubuh dan dapat berakibat fatal jika tidak ditangani. Prediksi penyakit jantung melalui pemodelan regresi dapat memungkinkan identifikasi dini yang lebih efektif. Penelitian ini bertujuan untuk memprediksi risiko penyakit jantung dengan menggunakan regresi linier sederhana berdasarkan data medis yang diperoleh dari data set yang tersedia di *Kaggle*. Analisis data dilakukan dengan *Python*, memanfaatkan *library* seperti *sklearn* untuk regresi, *matplotlib* dan *seaborn* untuk visualisasi, serta *numpy* dan *pandas* untuk manipulasi data. Data set yang digunakan terdiri dari 920 sampel dengan 16 atribut, di mana lima variabel independen dianalisis terhadap variabel. Hasil penelitian menunjukkan bahwa variabel *'Ca'* memiliki pengaruh paling signifikan terhadap risiko penyakit jantung, diikuti oleh *'Thalach'* dan *'Age'*. Sementara variabel *'Chol'* dan *'Thal'* memiliki pengaruh yang lebih kecil. Tujuan penelitian ini adalah menyediakan informasi prediktif yang akurat dan cepat, yang dapat membantu dalam pencegahan dan penanganan penyakit jantung. Temuan ini diharapkan dapat dijadikan acuan untuk strategi pencegahan dan pengambilan keputusan dalam bidang kesehatan.

Kata Kunci: *Linear Regression, Python, Heart Disease, Machine Learning, Google Colab*

Penulis Koresponden:
Dela Setiowati
Sistem Informasi, Rekayasa Industri, Universitas Telkom Purwokerto Jl. DI Panjaitan No.128, Karangreja, Purwokerto Kidul, Kec. Purwokerto Sel., Kabupaten Banyumas, Jawa Tengah 53147, Indonesia Email: delasetiowati@student.telkomuniversity.ac.id

1. PENDAHULUAN

Penyakit jantung merupakan tantangan yang mempengaruhi jutaan manusia di dunia yang menyebabkan kualitas hidup rendah dan masalah serius lainya pada tubuh manusia [1]. Jantung adalah organ tubuh manusia yang mempunyai fungsi memompa darah manusia ke seluruh tubuh pada manusia [2]. Penyakit pada jantung sangat mempengaruhi kesehatan pada tubuh manusia dan dapat menyebabkan kematian, jika tidak diatasi makan penyakit ini sangat fatal [3], penyakit jantung dapat menyerang siapa saja yang tidak bisa atau belum menerapkan hidup sehat. Gejala pada penyakit dan gangguan fungsi jantungnya sering kali tidak diketahui dan rasakan oleh para penderita, karena masyarakat telah kelalaian dan kurang dalam memperhatikan kesehatan jantungnya mereka [4]. Penyakit jantung adalah salah satu penyakit yang sering diderita masyarakat dengan gejala kelemahan fisik, kaki bengkak dan sesak nafas, dan menyerang siapa pun tanpa memandang usia, jenis kelamin, dan harapan hidup[5]. Menurut *World Health Organization* (WHO), pada tahun 2012, sebanyak 17,5 juta orang atau 31% dari populasi dunia meninggal akibat penyakit jantung, dan pada tahun 2018, tingkat kematian akibat penyakit kardiovaskular di Amerika

Serikat mencapai 217,1 per 100.000 penduduk [6]. Sistem kardiovaskular manusia merupakan salah satu sistem utama dalam tubuh manusia . Fungsi utama sistem ini adalah untuk menyebarkan beban ke seluruh jaringan. Tujuan sistem pernapasan adalah menyediakan nutrisi dan oksigen untuk tubuh manusia dan unggul dalam menghilangkan produk samping metabolisme. Salah satu organ penting dalam sistem kardiovaskular manusia adalah jantung. Jantung mempunyai tugas menghangatkan darah untuk memompa darah ke seluruh tubuh manusia. Jika jantung terkena dampak buruk atau terluka, hal ini akan berdampak negatif pada fungsi semua organ lain dalam tubuh manusia [7].

Penyakit jantung, yang sering kali disandingkan dengan istilah penyakit kardiovaskular atau angina, merupakan kondisi serius yang memerlukan perhatian dan penanganan yang tidak dapat diabaikan [8]. Gejala penyakit dan gangguan jantung lainnya sering kali tidak disadari oleh masyarakat atau penderitanya sendiri karena kurangnya perhatian atau kelalaian terhadap kesehatan jantungnya. Prediksi penyakit pada jantung dapat dilakukan dan dapat membantu untuk mencegah dan mengobati sebelum terlambat yang efektif. Pemodelan regresi dapat digunakan untuk mengetahui atau memprediksi sebuah penyakit pada jantung dengan menganalisis data medis [9]. Penggunaan *machine learning* pada dunia medis atau kesehatan sangat berkembang dengan pesat terutama dalam diagnosis yang di mana sebuah analisis dilakukan secara manual dapat dilakukan oleh analisis komputer [10]. Selanjutnya, dalam penelitian ini akan menjelaskan sebuah regresi data medis dapat dimanfaatkan untuk memperkirakan penyakit pada jantung dengan menganalisis data medis yang ada. Dengan menggunakan analisis regresi, peneliti dapat mengetahui, memahami dan mengidentifikasi pengaruh yang signifikan terhadap risiko penyakit jantung. Metode regresi ini sangat membantu apabila kita ingin menganalisis dan memprediksi kemungkinan kategori atau kejadian yang bergantung terhadap faktor-faktor lainnya. Metode regresi linier ini adalah salah satu metode statistik yang dimanfaatkan dalam memodelkan sebuah keterkaitan antara variabel independen atau independen dan variabel dependen atau yang bergantung[11].

Penelitian ini memiliki manfaat untuk mencari sebuah hasil prediksi yang nantinya bisa dijadikan sebagai acuan mengenai pencegahan penyakit jantung dan dapat dimanfaatkan untuk menentukan sebuah hasil yang baik dari dalam pengujian data tersebut. Hasil pengolahan data akan menjadi informasi dan pengetahuan yang diinginkan, sehingga dapat tergali potensi atau pengetahuan yang lebih berharga dan bermanfaat, hal ini akan menghasilkan analisis yang dapat memprediksi dan mencegah penyakit jantung serta menemukan peluang baru dan menciptakan analisis pada komputer untuk penyakit jantung. rencana untuk mencegah dan memprediksi penyakit jantung. Selain itu, dapat digunakan sebagai alat pengambilan keputusan.

2. METODE PENELITIAN

A. *Data Set Penelitian*

Data set adalah koleksi data yang dapat dipakai untuk eksperimen penelitian [12]. Istilah data set secara informal merujuk pada kumpulan data. Biasanya, data set terdiri dari beberapa variabel yang berhubungan dengan suatu topik spesifik. Data set juga dapat diartikan sebagai himpunan data dari informasi sebelumnya yang siap diolah menjadi informasi [13]. Pada penelitian ini data set yang digunakan diolah dari https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data . data set heart_disease_uci.csv berisi sebanyak 920 data yang mencangkup 16 kolom. Ke-16 kolom tersebut terdiri dari 8 kolom numerik (*id, age, trestbps, chol, thalch, oldpeak, ca, dan num) dan 8 kolom kategori (sex, dataset, cp, fbs, restecg, exang, slope, dan thal*). Selama proses analisis, data kategori akan diubah menjadi bentuk numerik. Pengubahan yang dijalankan *fixed defect* diganti dengan 0, normal diganti dengan 1, dan *reversable defect* diganti dengan 2.

B. *Eksplorasi Data*

Proses analisis data meliputi pengaturan dan pengelompokan data, serta pencarian pola atau tema guna memahami maknanya. Tahap analisis data dalam penelitian dilakukan setelah data yang dibutuhkan untuk menjawab permasalahan penelitian terkumpul dengan lengkap [14]. Langkah awal dalam analisis data adalah eksplorasi data, yang melibatkan penggunaan alat visualisasi dan teknik statistik untuk menemukan karakteristik dan pola awal dalam kumpulan data [15]. Memahami data melalui eksplorasi adalah langkah awal yang penting sebelum analisis dilakukan [16]. Eksplorasi data dilakukan untuk memahami data dengan lebih baik dan memperbaiki kualitas data yang dipakai. Tahap pertama dari proses ini adalah mengubah data mentah menjadi data yang dibutuhkan[17].

C. *Analisis Data*

Dalam tahap penelitian, analisis data diterapkan menggunakan regresi linier sederhana untuk memprediksi risiko penyakit jantung berdasarkan data medis dan diolah menggunakan Python [18]. *Google Collab* digunakan sebagai *platform* implementasi untuk mengintegrasikan analisis data. Data set yang digunakan berasal dari *website* www.kaggle.com yang awalnya berjumlah 16 kolom, yaitu *kolom id, age, sex, dataset, cp, trestbps, chol, fbs, restecg, thalch, exang, oldpeak, tilt, ca, thal, dan num.* Dalam penelitian ini, hanya lima kolom yang digunakan sebagai variabel independen, yaitu kolom *age, kol, thalch, ca, dan thal.* Sedangkan variabel yang dijadikan sebagai dependen teridkat adalah *num.*

D. *Diagram Alir Penelitian*

Diagram alir dalam penelitian ini adalah serangkaian urutan dari mulainya penelitian dilaksanakan sampai selesainya atau evaluasi dari penelitian yang dibuat, berikut adalah diagram alir penelitian pada Fig 1.

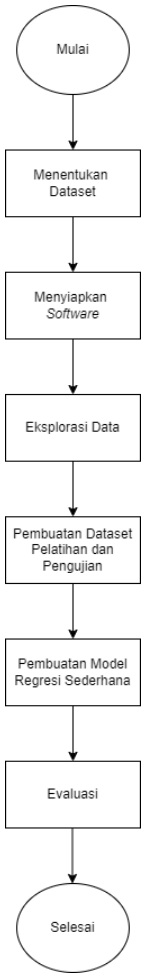

Fig. 1. Diagram Alir Penelitian

1. *Menentukan Data Set*

Pada tahap ini, penelitian dimulai dengan menentukan data set yang akan digunakan, sangat penting untuk analisis dan pencapaian tujuan penelitian. Proses ini melibatkan identifikasi kebutuhan data untuk memprediksi risiko penyakit jantung berdasarkan data medis pasien. Data set yang digunakan diambil dari *Kaggle*.

2. *Menyiapkan Software*

Pada langkah ini, menyiapkan perangkat lunak yang akan digunakan dalam penelitian dengan memilih *software* yang sesuai dengan jenis analisis yang akan dilakukan. Memastikan *software* yang dipilih dapat memenuhi kebutuhan analisis secara efektif dan efisien untuk mendukung hasil penelitian.

3. *Eksplorasi Data*

Proses eksplorasi data bertujuan untuk memahami karakteristik data serta memperbaiki kualitas data yang akan dipakai dalam analisis. Tahap ini melibatkan penyesuaian data mentah agar sesuai dengan kebutuhan analisis yang diinginkan.

4. *Pembuatan Data Set Pelatihan dan Pengujian*

Data set pelatihan digunakan untuk mengajarkan model cara mengenali pola dan menghasilkan prediksi. Data set ini berisi data yang telah dilabeli, memungkinkan model untuk belajar dari informasi tersebut. Setelah model selesai dilatih, data set pengujian digunakan untuk mengukur seberapa baik kinerjanya dengan data yang belum pernah diproses sebelumnya

5. *Pembuatan Model Regresi Sederhana*

Pembuatan model regresi yang lancar memudahkan proses pembuatan model matematika yang dapat memprediksi nilai dari satu variabel berdasarkan nilai variabel lain. Dalam regresi sederhana, variabel independen digunakan untuk memprediksi variabel dependen. Proses ini dimulai dengan mengumpulkan data yang relevan dan diakhiri dengan mengidentifikasi hubungan antara kedua variabel. Kemudian menggunakan teknik statistik untuk mengidentifikasi variabel yang menggambarkan hubungan ini, biasanya menggunakan ukuran sampel yang kecil. Grafik akan digunakan untuk memprediksi variabel dependen berdasarkan nilai variabel independen. Model regresi sederhana membantu untuk memahami dan memprediksi bagaimana perubahan pada satu variabel berpengaruh terhadap variabel lainnya.

6. *Evaluasi*

Evaluasi merupakan langkah yang penting dalam proses penelitian dan pengembangan suatu model, dengan tujuan mengidentifikasi beberapa model yang baik yang telah dikembangkan untuk klasifikasi data atau analisis regresi. Pada tahap ini, hasil model regresi sederhana yang dikembangkan akan dianalisis menggunakan data set pengujian.

3. HASIL DAN PEMBAHASAN

A. *Menyiapkan Software*

Software merupakan aplikasi yang terdiri dari serangkaian data elektronik yang diatur dan disimpan oleh perangkat komputer. Langkah pertama dalam analisis menggunakan *Python* adalah menyiapkan *library* yang akan dipakai. *Library* yang dipakai adalah *matplotlib*, *padas*, *pylab*, *numpy*, *sklearn*, *statsmodels*, dan *seaborn*. *Library numpy* berfungsi sebagai alat analisis data, sedangkan *library statsmodels* digunakan untuk analisis statistik dan eksplorasi data, *library matplotlib* dan *seaborn* membantu dalam visualisasi data, dan *library sklearn* dipakai untuk *machine learning*. Perintah untuk mengakses *library* pada Fig 2.

```
import matplotlib.pyplot as plt
import pandas as pd
import pylab as pl
import numpy as np
%matplotlib inline
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from sklearn.impute import SimpleImputer
import statsmodels.api as sm
from sklearn.metrics import r2_score, mean_squared_error
import seaborn as sns
```

Fig. 2. Library yang Digunakan

B. *Memanggil dan Menganalisis Data Set*

Sumber data set yang digunakan diambil dari www.kaggle.com . Untuk menggunakan data set, memakai fungsi *read* yang tersedia dalam *library pandas*. Pemanggilan dilakukan melalui perintah *df.head()*. Langkah ini untuk membaca data set awal. Hasil yang diperoleh dari perintah ada pada Fig 3.

```
df = pd.read_csv("/content/drive/MyDrive/TUBES AI BISMILLAH/heart_disease_uci.csv")
# melihat dataset
df.head()
```

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy	150.0	False	2.3	downsloping	0.0	fixed defect	0
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy	108.0	True	1.5	flat	3.0	normal	2
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy	129.0	True	2.6	flat	2.0	reversable defect	1
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal	187.0	False	3.5	downsloping	0.0	normal	0
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy	172.0	False	1.4	upsloping	0.0	normal	0

Fig. 3. Membaca Data Set

Pada Fig 3, data set heart_desease_uci.csv terdiri dari 16 kolom. Berdasarkan pada data set yang disebutkan, akan dibuat untuk melakukan prediksi terhadap risiko penyakit jantung yang dipengaruhi *age, chol, thalch, ca, dan thal.*

C. *Exploratory Data Analysis*

Exploratory Data Analysis membantu dalam menentukan teknik yang optimal untuk memanipulasi sumber data demi mendapatkan jawaban yang diinginkan, sehingga memudahkan data untuk menemukan pola, mengidentifikasi anomali, menguji hipotesis, atau menguji validitas asumsi [18]. Pada proses ini, melibatkan pemeriksaan informasi serupa dengan mencari informasi yang tidak ada, menghapus data duplikat, serta mengubah kategori menjadi nilai numerik. Hasil pengecekan informasi dataset ditunjukkan pada Fig 4 dan Fig 5.

```
df.describe()
```

	id	age	trestbps	chol	thalch	oldpeak	ca	num
count	920.000000	920.000000	861.000000	890.000000	865.000000	858.000000	309.000000	920.000000
mean	460.500000	53.510870	132.132404	199.130337	137.545665	0.878788	0.676375	0.995652
std	265.725422	9.424685	19.066070	110.780810	25.926276	1.091226	0.935653	1.142693
min	1.000000	28.000000	0.000000	0.000000	60.000000	-2.600000	0.000000	0.000000
25%	230.750000	47.000000	120.000000	175.000000	120.000000	0.000000	0.000000	0.000000
50%	460.500000	54.000000	130.000000	223.000000	140.000000	0.500000	0.000000	1.000000
75%	690.250000	60.000000	140.000000	268.000000	157.000000	1.500000	1.000000	2.000000
max	920.000000	77.000000	200.000000	603.000000	202.000000	6.200000	3.000000	4.000000

Fig. 4. Deskripsi Data

Fig 4 adalah analisis untuk memeriksa data statistik. Data tersebut dapat diterapkan untuk memeriksa adanya data yang tidak sesuai. Dari data itu, dapat dianalisis nilai-nilai tertentu, misalnya age tertinggi adalah 77. Usia maksimum 77 masih dianggap wajar.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 920 entries, 0 to 919
Data columns (total 16 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        920 non-null    int64
 1   age       920 non-null    int64
 2   sex       920 non-null    object
 3   dataset   920 non-null    object
 4   cp        920 non-null    object
 5   trestbps  861 non-null    float64
 6   chol      890 non-null    float64
 7   fbs       830 non-null    object
 8   restecg   918 non-null    object
 9   thalch    865 non-null    float64
 10  exang     865 non-null    object
 11  oldpeak   858 non-null    float64
 12  slope     611 non-null    object
 13  ca        309 non-null    float64
 14  thal      434 non-null    object
 15  num       920 non-null    int64
dtypes: float64(5), int64(3), object(8)
memory usage: 115.1+ KB
```

Fig. 5. Informasi Data Set

Pada Fig 5, diperlihatkan informasi mengenai data set yang terdiri dari 920 dan 16 kolom. Dalam Fig 4, terlihat bahwa tidak tersedia data yang kosong dan setiap kolom memiliki tipe data yang jelas. Pada kolom *id, age, trestbps, chol, thalch, oldpeak, ca,* dan *num* memiliki tipe numerik (*integer dan float*) dan kolom *sex, cp, fbs, restecg, exang, slope,* dan *thal* bertipe *object*. Untuk mempermudah ketika melakukan analisis data maka, data yang bertipe objek akan diimplementasikan *encoding* agar menjadi numerik. Proses *encoding* dapat dilihat pada Fig 6

```
num_cols =df.select_dtypes(include=np.number).columns
num_cols
non_num_cols =df.select_dtypes(exclude=np.number).columns
non_num_cols
label_encoder =LabelEncoder()
for i in non_num_cols:
    df[i]=label_encoder.fit_transform(df[i])
df.head(10)
```

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1	63	1	0	3	145.0	233.0	1	0	150.0	0	2.3	0	0.0	0	0
1	2	67	1	0	0	160.0	286.0	0	0	108.0	1	1.5	1	3.0	1	2
2	3	67	1	0	0	120.0	229.0	0	0	129.0	1	2.6	1	2.0	2	1
3	4	37	1	0	2	130.0	250.0	0	1	187.0	0	3.5	0	0.0	1	0
4	5	41	0	0	1	130.0	204.0	0	0	172.0	0	1.4	2	0.0	1	0
5	6	56	1	0	1	120.0	236.0	0	1	178.0	0	0.8	2	0.0	1	0
6	7	62	0	0	0	140.0	268.0	0	0	160.0	0	3.6	0	2.0	1	3
7	8	57	0	0	0	120.0	354.0	0	1	163.0	1	0.6	2	0.0	1	0
8	9	63	1	0	0	130.0	254.0	0	0	147.0	0	1.4	1	1.0	2	2
9	10	53	1	0	0	140.0	203.0	1	0	155.0	1	3.1	0	0.0	2	1

Fig. 6. Hasil *Encoding*

Hasil yang ditunjukkan pada Fig 6 mengungkapkan bahwa tipe data pada berbagai kolom seperti *sex, cp, fbs, restecg, exang, slope,* dan *thal* telah berhasil diubah menjadi tipe data numerik.

D. *Analisis Regresi Linier*

Regresi linier merupakan metode statistik yang diterapkan dalam menganalisis pengaruh satu atau lebih variabel terhadap variabel lainnya .Penggunaan *library sklearn* digunakan untuk melakukan perhitungan regresi linier. Berikut ini merupakan proses analisis regresi linier dalam perintah *Python*.

1. *Menentukan Kolom Variabel Independen dan Variabel Dependen*

```
[ ]  x = df[['age','chol','thalch','ca','thal']]
     y = df['num']
```

Fig. 7. Menentukan Kolom

Kode tersebut digunakan untuk memisahkan data menjadi variabel independen (x) dan variabel dependen (y) yang akan dipakai dalam analisis atau pemodelan. Variabel independent terdiri dari kolom *age, chol, thalch, ca, dan thal*, sedangkan variabel dependent terdiri dari kolom *num*.

2. *Menentukan dan Membagi Data Menjadi Data Pelatihan dan Data Pengujian*

```
[ ]  x_train, x_test, y_train, y_test = train_test_split(x, y,
     test_size=0.2, random_state=4)
```

Fig. 8. Membagi Data

Kode tersebut membagi data set menjadi set pelatihan dan set pengujian melalui fungsi *train_test_split dari scikit-learn.* Sebanyak 20% data dialokasikan untuk set pengujian dan 80% untuk set pelatihan.

3. *Menghapus Baris yang Mengandung Nilai yang Hilang*

```
[ ]  train_data = pd.concat([x_train, y_train], axis=1)

[ ]  train_data_clean = train_data.dropna()
```

Fig. 9. Menghapus Baris dengan Nilai Hilang

Kode ini menggabungkan *x_train* dan *y_train* menjadi *DataFrame train_data*, kemudian membersihkannya dari nilai yang hilang dan menghasilkan *train_data_clean.* Langkah ini memastikan data pelatihan bebas dari nilai kosong untuk meningkatkan kualitas model.

4. *Melatih Model Regresi Linier*

```
[ ]  lin_reg = LinearRegression()
     lin_reg.fit(x_train_clean, y_train_clean)
```

```
▾ LinearRegression
LinearRegression()
```

Fig. 10. Melatih Model Regresi Linier

Proses ini menciptakan model regresi linier dan melatihnya menggunakan data pelatihan yang telah dibersihkan.

5. *Menampilkan Nilai Variabel Independen*

```
[ ]  feature_cols = ['age','chol','thalch','ca','thal']
     X = df[feature_cols]
     y = df['num']
     list(zip(feature_cols, lin_reg.coef_))
```

```
[('age', 0.000658599901168035),
 ('chol', -0.005365507423581307),
 ('thalch', 0.3248818952109049),
 ('ca', -0.34611252324452424),
 ('thal', -0.24167594632520756)]
```

Fig. 11. Menampilkan Nilai Variabel Independen

Langkah ini mendefinisikan fitur yang digunakan dalam model regresi linier dan mengekstrak data dari data set. Variabel fitur berisi informasi seperti *age, cholc, thalch, ca*, dan *thal*, sementara variabel target berisi data yang ingin diprediksi.

E. *Nilai Korelasi*

Nilai korelasi digunakan untuk mengeksplorasi hubungan antara dua variabel [21]. Untuk melihat dan mengukur sejauh mana hubungan masing-masing variabel independen terhadap variabel dependen sehingga korelasi persial perlu dihitung. dapat dilihat pada fig 12 di bawah ini.

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	nu
id	1.000000	0.239301	0.280053	0.949062	-0.189430	0.052924	-0.376936	0.291005	0.477040	-0.466427	0.399679	0.049930	0.115371	0.061433	0.484673	0.27355
age	0.239301	1.000000	0.056889	0.235076	-0.076519	0.244253	-0.086234	0.125887	-0.013094	-0.365778	0.250938	0.258243	-0.222399	0.370416	-0.055761	0.33959
sex	0.280053	0.056889	1.000000	0.285734	-0.125933	0.001087	-0.197281	0.106900	0.074900	-0.179320	0.207197	0.103930	-0.033180	0.094123	0.148581	0.25934
dataset	0.949062	0.235076	0.285734	1.000000	-0.150334	0.021227	-0.416648	0.293678	0.456794	-0.414609	0.399767	0.053002	0.131545	0.030384	0.464508	0.27620
cp	-0.189430	-0.076519	-0.125933	-0.150334	1.000000	-0.023508	0.065279	-0.078563	-0.064280	0.300812	-0.241050	-0.181486	0.183725	-0.199452	-0.096112	-0.31451
restbps	0.052924	0.244253	0.001087	0.021227	-0.023508	1.000000	0.092853	-0.011590	0.014034	-0.104899	0.152328	0.161908	-0.029471	0.093705	-0.004139	0.12229
chol	-0.376936	-0.086234	-0.197281	-0.416648	0.065279	0.092853	1.000000	-0.412369	-0.202552	0.236121	-0.029707	0.047734	0.087924	0.051606	-0.041309	-0.23154
fbs	0.291005	0.125887	0.106900	0.293678	-0.078563	-0.011590	-0.412369	1.000000	0.119873	-0.135389	0.004113	0.006027	-0.063412	0.193544	0.071754	0.18666
restecg	0.477040	-0.013094	0.074900	0.456794	-0.064280	0.014034	-0.202552	0.119873	1.000000	-0.170422	0.141274	-0.037788	0.171056	-0.114783	0.336679	0.03425
thalch	-0.466427	-0.365778	-0.179320	-0.414609	0.300812	-0.104899	0.236121	-0.135389	-0.170422	1.000000	-0.356439	-0.151174	0.170790	-0.264094	-0.173028	-0.36626
exang	0.399679	0.250938	0.207197	0.399767	-0.241050	0.152328	-0.029707	0.004113	0.141274	-0.356439	1.000000	0.393714	-0.108577	0.127385	0.184886	0.33816
oldpeak	0.049930	0.258243	0.103930	0.053002	-0.181486	0.161908	0.047734	0.006027	-0.037788	-0.151174	0.393714	1.000000	-0.589337	0.281817	0.027792	0.44308
slope	0.115371	-0.222399	-0.033180	0.131545	0.183725	-0.029471	0.087924	-0.063412	0.171056	0.170790	-0.108577	-0.589337	1.000000	-0.121233	0.228077	-0.31838
ca	0.061433	0.370416	0.094123	0.030384	-0.199452	0.093705	0.051606	0.193544	-0.114783	-0.264094	0.127385	0.281817	-0.121233	1.000000	0.136697	0.51621
thal	0.484673	-0.055761	0.148581	0.464508	-0.096112	-0.004139	-0.041309	0.071754	0.336679	-0.173028	0.184886	0.027792	0.228077	0.136697	1.000000	-0.00517
num	0.273552	0.339596	0.259342	0.276203	-0.314518	0.122291	-0.231547	0.186664	0.034255	-0.366265	0.338166	0.443084	-0.318383	0.516216	-0.005170	1.00000

Fig. 12. Nilai Korelasi

Langkah ini digunakan untuk menghitung dan menampilkan matriks korelasi antara semua kolom dalam *DataFrame*. Korelasi menunjukkan sejauh mana dua variabel berhubungan satu sama lain. Berikut adalah nilai korelasi antara variabel independen dengan dependen.

Table. 1 Korelasi Variabel Independen dengan Dependen

	Age	*Chol*	*Thalach*	*Ca*	*Thal*	*Num*
Age	1.000000	-0.08	-0.36	0.37	-0.05	0.33
Chol	-0.08	1.000000	0.23	0.05	-0.04	-0.23
Thalach	-0.36	0.23	1.000000	-0.26	-0.17	-0.36
Ca	0.37	0.05	-0.26	1.000000	0.13	0.51
Thal	-0.05	-0.04	-0.17	0.13	1.000000	-0.005

Koefisien korelasi, yang berkisar antara -1 hingga +1, menunjukkan derajat hubungan antara dua variabel, di mana nilai mendekati +1 mencerminkan hubungan positif yang erat, mendekati -1 mencerminkan hubungan negatif yang erat, dan mendekati 0 mencerminkan hubungan yang lemah atau tidak signifikan. Hasil korelasi antar data pada tabel 1, menunjukkan ada hubungan positif cukup kuat antara '*Ca'* dengan Num (0.51), hubungan negatif sedang antara '*Thalach'* dengan *'Num'* (-0.36), dan hubungan positif sedang antara *'Age'* dengan *'Num'* (0.33). Hal ini bisa diprediksi orang dengan jumlah pembuluh darah utama yang mengalami kelainan lebih banyak (*'Ca'* lebih tinggi) memiliki risiko lebih besar terkena penyakit jantung serius. Prediksi dari korelasi antara *'Thalach'* dan *'Age'* dengan *'Num'*, semakin tinggi detak jantung maksimum yang tercapai dan semakin tinggi usia seseorang, maka semakin tinggi pula risiko terkena penyakit jantung.

F. *Uji Koefisien Determinasi*

```
ypredict = lin_reg.predict(x_test)
```

```
print ("Coefficient of determination :",r2_score(y_test,ypredict))
print ("MSE: ",mean_squared_error(y_test,ypredict))
print("RMSE: ",np.sqrt(mean_squared_error(y_test,ypredict)))
```

Fig. 13. Menguji Koefisien

Koefisien determinasi ini digunakan untuk mengukur sejauh mana pengaruh variabel independen atau dependen terhadap variabel dependen yang selanjutnya menentukan persamaan dan kesesuaian model regresi linier [2]. hasil perhitungan dapat dilihat pada tabel berikut

Table. 2. NILAI KORELASI

Nilai Korelasi	
Coefficient of determination	0.24463199862495943
MSE	1.1037726492517173
RMSE	1.0506058486662433

G. *Perbandingan Nilai Aktual dan Nilai Prediksi*

```
df_best_predict = pd.DataFrame({'Actual': y_test, 'Predicted': ypredict})
df_best_predict.head(10)
```

Fig. 14 Perbandingan Nilai Aktual dan Prediksi

Hasil dari perbandingan tersebut dapat dilihat pada tabel 2.

Table. 3. PERBANDINGAN NILAI

Actual	Predicted
291	0.238439
9	0.698151
57	0.742926
60	1.049002
25	0.264649
63	0.139776
92	2.592393
185	0.977899
246	1.272446
46	0.744692

Langkah ini menciptakan *DataFrame* baru yang berisi nilai aktual dan nilai prediksi dari model. Kolom pertama menunjukkan nilai sebenarnya, sedangkan kolom kedua menampilkan hasil prediksi. Menampilkan sepuluh baris pertama membantu dalam mengevaluasi kinerja model dalam menghasilkan prediksi yang akurat.

H. *Visualisasi Perbandingan Nilai*

```
plt.figure(figsize=(10,7))
plt.title("Actual vs. predicted",fontsize=10)
plt.xlabel("Actual",fontsize=10)
plt.ylabel("Predicted", fontsize=10)
#plt.scatter(x=test_y,y=test_predict)
sns.regplot(x=y_test, y=ypredict)
plt.show()
```

Fig. 15. Menampilkan Perbandingan dalam Grafik

Langkah ini digunakan untuk memvisualisasikan perbandingan antara nilai sebenarnya dan nilai yang diprediksi oleh model regresi linier. Plot ini menggunakan *scatter plot* dan garis regresi untuk menunjukkan hubungan antara nilai aktual dan prediksi, sehingga membantu dalam mengevaluasi kinerja model secara visual. Berikut ini adalah visualisasi grafik yang dihasilkan:

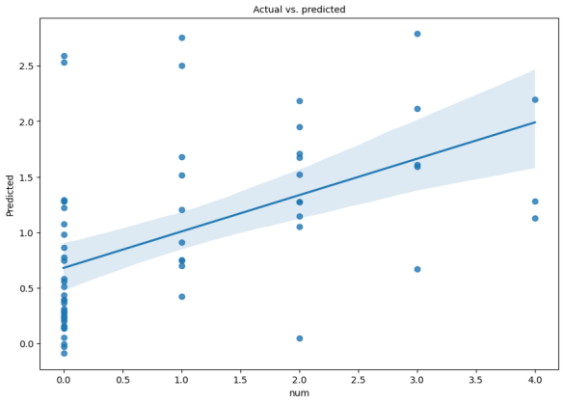

Fig. 16. Hasil dalam Bentuk Grafik

Hasil pada tabel di atasterdapat perbedaan antara data Y aktual dan Y prediksi, yang kemudian ditampilkan dalam bentuk grafik pada fig diatas.

I. *Visualisasi Pengaruh Variabel Independen terhadap Variabel Dependen*

```
[ ] plt.figure(figsize=(10, 6))
    plt.bar(df_coef_sorted['Feature'], df_coef_sorted['Coefficient'], color='skyblue')
    plt.xlabel('Variabel Bebas')
    plt.ylabel('Koefisien')
    plt.title('Pengaruh Variabel Bebas terhadap Num')
    plt.xticks(rotation=45)
    plt.grid(axis='y', linestyle='--', alpha=0.7)

    for index, value in enumerate(df_coef_sorted['Coefficient']):
        plt.text(index, value + 0.01 if value > 0 else value - 0.04, f'{value:.3f}', ha='center', va='bottom', fontsize=9)

    plt.tight_layout()
    plt.show()
```

Fig. 17. Membuat Diagram

Langkah ini digunakan untuk membuat visualisasi grafik batang yang menunjukkan pengaruh variabel bebas terhadap variabel target (*num*). Dengan ukuran grafik yang ditentukan, grafik ini menampilkan nama variabel bebas di sumbu x dan nilai koefisien di sumbu y, yang merepresentasikan kontribusi masing-masing variabel terhadap prediksi model. Berikut ini adalah peringkat pengaruh risiko penyakit jantung berdasarkan hasil prediksi.

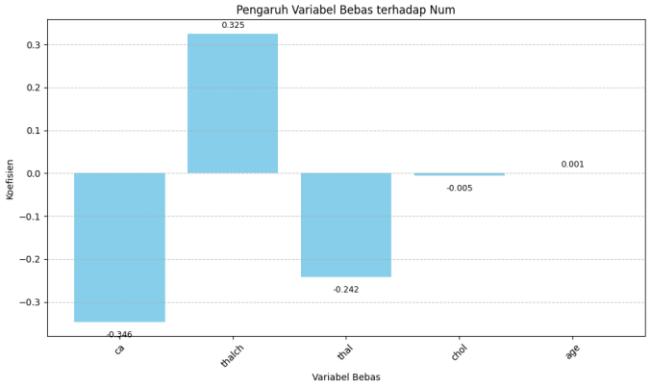

Fig. 18. Peringkat Pengaruh Risiko

4. KESIMPULAN

Berdasarkan analisis koefisien dari model regresi linier yang telah dilakukan, dapat disimpulkan bahwa variabel *'Ca'* (jumlah pembuluh darah utama yang mengalami kelainan) memiliki pengaruh paling signifikan terhadap variabel target *'Num'*, yang merupakan indikasi penyakit jantung, dengan koefisien positif sebesar 0.51. Hal ini menunjukkan bahwa semakin tinggi nilai *'Ca'*, semakin tinggi pula nilai *'Num'*, yang berarti risiko penyakit jantung meningkat. Variabel *'Thalach'* (detak jantung maksimum) menunjukkan pengaruh yang cukup signifikan dengan koefisien negatif -0.36, mengindikasikan bahwa semakin tinggi nilai *'Thalach'*, semakin rendah risiko seseorang terkena penyakit jantung, yang menunjukkan adanya hubungan negatif sedang antara detak jantung maksimum yang tercapai dengan risiko penyakit jantung. Selanjutnya, variabel *'Age'* (usia) memiliki koefisien positif sebesar 0.33, yang berarti semakin tinggi nilai *'Age'*, semakin tinggi nilai *'Num'*, menunjukkan peningkatan risiko penyakit jantung. Variabel *'Chol'* (kolesterol) memiliki pengaruh yang lebih kecil terhadap *'Num'* dengan koefisien negatif sebesar -0.23, menunjukkan bahwa peningkatan *'Chol'* sedikit menurunkan nilai *'Num'* dan risiko penyakit jantung. Terakhir, variabel *'Thal'* (*Thalassemia*) memiliki koefisien negatif yang sangat kecil sebesar -0,005, yang menunjukkan bahwa pengaruh *Thalassemia* terhadap *'Num'* sangat kecil dan hampir tidak terlihat dalam model. Dengan demikian, dapat disimpulkan bahwa pada model regresi linier yang diterapkan, *'Ca'* memiliki pengaruh paling signifikan terhadap risiko penyakit jantung, diikuti oleh *Thalach'* dan *'Age'*, sedangkan *'Chol'* dan *'Thal'* memiliki pengaruh yang lebih kecil.

UCAPAN TERIMAKASIH

Kami mengucapkan terima kasih kepada pembimbing atas arahan, dukungan, serta masukan yang sangat berarti selama proses penyusunan jurnal ini. Ucapan terima kasih juga kami sampaikan kepada semua pihak yang telah memberikan bantuan dan dukungan, baik secara langsung maupun tidak langsung. Kami juga mengapresiasi kerja sama tim yang baik dan penuh tanggung jawab, yang telah mendukung kelancaran penelitian ini. Kami berharap penelitian ini dapat memberikan manfaat dan kontribusi positif bagi kemajuan ilmu pengetahuan.

DAFTAR PUSTAKA

[1] E. I. Scandea, M. Aqsha, R. Sugiarto, F. Lestari, And D. Hartanti, "Penerapan Data Mining Untuk Menganalisis Data Faktor Resiko Penyakit Jantung Menggunakan Metode Logistic Regression," Pp. 683–688, 2023.

[2] D. K. Saputro, M. Fiko, R. Ajie, S. Azizah, And D. Hartanti, "Penerapan Logistic Regression Untuk Mendeteksi Penyakit Jantung Pada Pasien," Pp. 666–671, 2023.

[3] U. Amelia, J. Indra, And A. F. N. Masruriyah, "Implementasi Algoritma Support Vector Machine (Svm) Untuk Prediksi Penyakit Stroke Dengan Atribut Berpengaruh," *Sci. Student J. Information, Technol. Sci.*, Vol. Iii, No. 2, Pp. 254–259, 2022.

[4] K. Kevin, "Diagnosa Penyakit Jantung Menggunakan Metode Certainty Factor," *J. Inform. Dan Rekayasa Perangkat Lunak*, Vol. 3, No. 1, Pp. 93–106, 2022.

[5] A. Khan And A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification In E-Healthcare," Vol. 8, No. Ml, 2020.

[6] A. Muliawan, A. Rizal, S. Hadiyoso, And C. Author, "Heart Disease Prediction Based On

Physiological Parameters Using Ensemble Classifier And Parameter," Vol. 5, No. 1, Pp. 258–267, 2023.

[7] A. B. Wibisono And A. Fahrurozi, "Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner," *J. Ilm. Teknol. Dan Rekayasa*, Vol. 24, No. 3, Pp. 161–170, 2019.

[8] D. Liegar, S. I. Junaidi, And S. M. Isa, "International Journal Of Advanced Trends In Computer Science And Engineering Available Online At Http://Www.Warse.Org/Ijatcse/Static/Pdf/File/Ijatcse146922020.Pdf Artificial Neural Network Architecture Optimization For Heart Disease," 2020.

[9] A. Wijayadhi, M. Makmun Effendi, And S. Budi Rahardjo, "Prediksi Penyakit Jantung Dengan Algoritma Regresi Linier," *Bull. Inf. Technol.*, Vol. 4, No. 1, Pp. 15–28, 2023.

[10] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, And R. S. Suraj, "Heart Disease Prediction Using Hybrid Machine Learning Model," 2021.

[11] A. S. Novari And U. K. Nisak S, "Prediksi Faktor Yang Mempengaruhi Hipertensi Dengan Metode Data Mining Untuk Meningkatkan Pelayanan Kesehatan Di Upt Puskesmas Ngoro," *Phys. Sci. Life Sci. And Engineering*, Vol. 1, No. 2, P. 16, 2024.

[12] Ander Sriwi Sri Sucaty, Murianto, "Penerapan Algoritma Artificial Neural Network Untuk Memprediksi Penyakit Gagal Jantung," *Pola Kemitraan Pentahelix Dalam Pengemb. Desa Wisata Buwun Sejati, Lomb. Barat Ntb*, Vol. 3, No. 4, Pp. 413–446, 2024.

[13] S. P. Tamba And E. -, "Prediksi Penyakit Gagal Jantung Dengan Menggunakan Random Forest," *J. Sist. Inf. Dan Ilmu Komput. Prima(Jusikom Prima)*, Vol. 5, No. 2, Pp. 176–181, 2022.

[14] S. R. Ivan Fanami Qomaruddin, "Analisis Data Kuantitatif Dengan Program Ibm Spss." P. Hal 1, 2021.

[15] C. Raras, A. Widiawati, L. Nurazizah, And I. R. Yunita, "Implementasi Algoritma Logistic Regression Pada Pembuatan Website Sederhana Untuk Prediksi Penyakit Jantung," *Infotekmesin*, Vol. 15, No. 01, Pp. 117–122, 2024.

[16] T. A. F. M. Dan I. P. A. U. Team, "Analisis Data Eksplorasi," *Transformasi Data*. 2018.

[17] M. Sholeh, S. Suraya, And D. Andayati, "Machine Linear Untuk Analisis Regresi Linier Biaya Asuransi Kesehatan Dengan Menggunakan Python Jupyter Notebook," *J. Edukasi Dan Penelit. Inform.*, Vol. 8, No. 1, P. 20, 2022.

[18] A. N. Hanna, J. S. Mcdonald, C. H. Miller, And D. Couri, "Pretreatment With Paracetamol Inhibits Metabolism Of Enflurane In Rats," *Br. J. Anaesth.*, Vol. 62, No. 4, Pp. 429–433, 1989.