Perbandingan Algoritma *Logistic Regression* dengan *Decision Tree* untuk Memprediksi Penyakit Jantung Menggunakan Bahasa Pemrograman Pyhton

Reinhard[1], Katherine Allen Lius[2], Theresia Cindana[3], Tralya Dharmada[4], Valencia Eurelia A.T[5]

[1] reinhard@student.umn.ac.id, [2] katherine.lius@student.umn.ac.id, [3] theresia.cindanda@student.umn.ac.id,

[4] tralya.dharmada@student.umn.ac.id, [5] valencia.tania@student.umn.ac.id

Program Studi Sistem Informatika, Fakultas Teknik dan Informatika,

Universitas Multimedia Nusantara

***Abstract* —** Heart disease is one of the most deadly and life-threatening chronic diseases. It is one of the most common causes of mortality in different areas of the world each year. From this study, we can compare the prediction result using two different algorithms. The Analysis result that are being compared are Logistic Regression and Decision Tree Algorithm. The results of the study of heart disease predictions by comparing Logistic Regression Algorithm and Decision Tree Algorithm shows the highest accuracy value of 85% using the Logistic Regression algorithm. So, we can conclude that Logistic Regression is the better algorithm to predicting early heart disease in patients heart symptoms.

***Index Terms* — Classification, Decision Tree, Heart Disease, Logistic Regression, Machine Learning, Prediction, Phyton, CRISP-DM**

I. PENDAHULUAN

Jantung merupakan organ paling penting dalam tubuh manusia. Peran jantung sangat mempengaruhi organ lain karena berfungsi untuk memompa darah yang mengandung oksigen serta nutrisi ke seluruh tubuh. Pentingnya peranan jantung menjadi salah satu alasan penyakit ini sangat ditakuti oleh masyarakat. Mengingat, penyakit ini menjadi salah satu jenis penyakit yang sangat berbahaya dan mengancam keselamatan. Pada tahun 2019, berdasarkan data dari WHO, 32% kematian didunia disebabkan oleh kardiovaskular (CVDs) dan telah merenggut sebanyak 17,9 juta korban jiwa[1].

Selama 50 tahun terakhir, jumlah orang yang menderita penyakit jantung koroner semakin meningkat tanpa mengetahui alasan yang jelas dari penyakit tersebut[2]. Namun dewasa ini, banyak penelitian yang menyatakan bahwa penyakit jantung memiliki banyak sekali penyebab. Menurut Arif Muttaqin, penyebab penyakit jantung dibedakan menjadi dua macam faktor yaitu dapat diubah dan tidak dapat diubah. Faktor yang tidak dapat diubah mencakup usia dan jenis kelamin. Sedangkan, faktor yang dapat diubah seperti hipertensi dan kadar kolesterol. Pasien dengan riwayat hipertensi memiliki peluang penyakit jantung lebih tinggi dibandingkan orang dengan tekanan darah normal. Selain itu, kadar kolesterol dapat memicu penyakit jantung dikarenakan timbunan kolesterol dalam pembuluh darah dapat menghambat aliran oksigen masuk ke jantung untuk dipompa ke seluruh tubuh[3].

Selain itu, ada juga beberapa faktor lain yang mempengaruhi penyakit jantung, seperti detak jantung yang tidak stabil. Melansir dari laman halodoc.com, detak jantung normal orang dewasa berada pada rentang 60-100 kali per menit[4]. Selain itu, aktivitas fisik juga berhubungan langsung dengan detak jantung seseorang. Berdasarkan penelitian Silvia, seseorang yang memiliki aktivitas ringan memiliki risiko lebih besar terkena penyakit jantung koronor dari pada orang dengan aktivitas sedang dan berat.[5] Tetapi bukan berarti orang dengan aktivitas berat terhindar dari penyakit jantung karena penelitian menunjukkan bahwa penyebab medis paling tinggi seorang atlet meninggal dikarenakan serangan henti jantung mendadak (*sudden cardiac death*/ SCD)[6].

Penyakit jantung tidak hanya berbahaya, tetapi juga sulit untuk diprediksi. Kesulitan ini juga ditambah

[1] WHO. Cardiovascular diseases (CVDs). June 11, 2021. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [Accessed 9 October 2021].

[2] Shoufiah, R. *Hubungan Faktor Resiko dan Karakteristik Penderita dengan Kejadian Penyakit Jantung Koroner*. Mahakam Nursing Journal Vol 1, No. 1, May 2016 : 17-26

[3] R. E. H. Patriyani and D. F. Purwanto, "Faktor Dominan Risiko Terjadinya Penyakit Jantung (PJK)," *Keperawatan Global*, vol. 1, no. 1, p. 24, 2016.

[4] Halodoc, R. (2021, November 25). Detak jantung normal seseorang ditentukan berdasarkan usia hingga aktivitas yang dilakukan. Simak selengkapnya. halodoc. https://www.halodoc.com/artikel/berapa-detak-jantung-normal-berdasarkan-usia

[5] I. Ramadini and S. Lestari, "Hubungan Aktivitas Fisik dan Stress dengan Nyeri Dada Pasien Penyakit Jantung Koroner," *Human Care*, vol. 2, no. 3, 2017.

[6] -. (2011, September 6). Kematian Akibat Serangan Jantung Banyak Menimpa Atlet Muda. detikHealth. https://health.detik.com/berita-detikhealth/d-1716590/kematian-akibat-serangan-jantung-banyak-menimpa-atlet-muda

karena minimnya jumlah dokter spesialis jantung. Oleh karena itu, keberadaan rekam medis digital pasien dapat sangat membantu bidang ilmu kedokteran dalam hal memprediksi penyakit jantung. Rekam digital tersebut kemudian dapat dimanfaatkan bersamaan dengan *machine learning*. Berbekal kedua hal tersebut, sebuah informasi dapat diperoleh untuk memprediksi penyakit jantung.

Machine learning sendiri adalah sebuah metode analisa yang digunakan untuk menyelesaikan persoalan yang rumit dan berskala besar, salah satunya mengubah rekam medis menjadi pengetahuan[7]. Adanya teknologi *machine learning* dapat membantu manusia mengambil keputusan dengan berbagai pertimbangan dan melakukan prediksi masa depan.

Penelitian ini menggunakan algoritma Logistic Regresion dan Decision Tree untuk membuat klasifikasi terkait penyakit jantung berdasarkan variabel-variabel yang ada. Selain itu, tujuan digunakannya dua algoritma adalah untuk membandingkan model mana yang lebih akurat dalam memprediksi penyakit jantung. Adapun, permasalahan yang diangkat dalam penelitian ini adalah sebagai berikut:

1. Bagaimana hasil yang diperoleh dalam memprediksi penyakit jantung, menggunakan algoritma Logistic Regression dan Decision Tree pada bahasa pemrograman Python?
2. Apa algoritma yang lebih baik digunakan untuk memprediksi penyakit jantung?

Berdasarkan permasalahan tersbut, tujuan dilakukannya penelitian ini adalah:

1. Mengetahui hasil prediksi dari kedua algoritma untuk memprediksi penyakit jantung.
2. Mencari algoritma yang memiliki tingkat akurasi paling tinggi dalam memprediksi penyakit jantung.

II. TINJAUAN PUSTAKA

A. *Machine Learning*

Machine learning merupakan sebuah ilmu yang menggunakan teknologi pembelajaran mesin secara matematis yang banyak digunakan untuk menyelesaikan suatu permasalahan dan memudahkan pengerjaan suatu kegiatan[8]. Penggunaan *machine learning* banyak digunakan untuk mendapatkan suatu informasi atau wawasan dari sekumpulan data yang dimiliki agar informasi tersebut dapat digunakan secara lebih efisien atau efektif, misalnya untuk melihat suatu pola, memprediksi suatu kejadian, dan membantu manusia dalam menentukan keputusan.

B. *Logistic Regression*

Logistic regression adalah sebuah algotirma analitik yang biasanya digunakan untuk mempelajari hubungan antara variabel dependen yang bersifat biner dan variabel independen yang bersifat numerik atau kategorik[9]. Adapun, variabel biner adalah sebuah data bersifat kategorikal yang hanya memiliki 2 (dua) nilai nominal atau ordinal yang digunakan untuk memprediksi (0 dan 1). Dengan mempelajar hubungan yang tercipta antar variabel, algoritma *logistic regression* dapat digunakan untuk melakukan prediksi masa depan.

Logistic regression dibagi lagi menjadi dua macam yaitu *simple logistic regression* dan *multiple logistic regression*. Sesuai dengan namanya, *simple logistic regression* adalah sebuah algoritma yang hanya menganalisa hubungan sederhana yang terjalin antara satu variabel dependen dengan satu variabel independen. Sedangkan di lain sisi, *multiple logistic regression* adalah sebuah algoritma yang digunakan untuk mencari hubungan antara satu variabel dependen dengan banyak variabel independen. Dalam situasi ini, model harus berkorelasi secara liniar dengan peluang log dan memiliki multikolinearitas atau tidak sama sekali[10].

Sebagai sebuah algoritma, *logistic regression* memiliki kelebihan serta kekurangan[11]. Adapun kekurangan dari algoritma ini yaitu:

1. Prediksi menggunakan algoritma ini hanya dapat berlangsung jika tidak ditemukan multikolinearitas pada variabel-variabel independennya.
2. Untuk dataset dengan kelas yang tidak seimbang, algoritma ini rentan mengalami *underfitting* sehingga akurasi yang dihasilkan rendah[12].

Di lain sisi, algoritma *logistic regression* juga memiliki keunggulan seperti:

---

[7] Escamila, A.K. Hassani, A. H & Andres, E. *Classification models for heart disease prediction using feature selection and PCA*. Volume 19, 2020.

[8] Telaumbanua, F. D., Hulu, P., Nadeak, T. Z., Lumbantong, R. R., & Dharma, A.. *Penggunaan Machine Learning Di Bidang Kesehatan.* JURNAL TEKNOLOGI DAN ILMU KOMPUTER PRIMA (JUTIKOMP), 2(2), 391-399. 2020. https://doi.org/10.34012/jutikomp.v2i2.657

[9] Nurdiansah, S.N & Khikmah, Laelatul. *Binary Logistic Regression Analysis of Variables that Influence Poverty in Central Java*. Vol. 1, No. 1, March 2020

[10] E. Y. Boateng, and D. A. Abaye, "*A Review of the Logistic Regression Model with Emphasis on Medical Research"*. Journal of Data Analysis and Information Processing, Vol.7. No.4, pp.190-207, 2019.

[11] Edgar, T.W. & Manz, D.O. *Research Methods for Cyber Security.* 2017

[12] Rianto, H. Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software. Journal of Software Engineering, Vol. 1, No. 1, April 2015

1. Dapat memberikan ukuran prediktor yang tepat.
2. Algoritma ini dapat melihat arah hubungan yang tercipta (positif atau negatif).
3. *Logistic regression* menghasilkan variabel yang dapat dipisahkan secara linear dan mudah untuk diterapkan, dianalisa serta diuji coba.

C. *Decision Tree*

Decision tree atau pohon keputusan merupakan sebuah algoritma klasifikasi yang berbentuk seperti pohon. Algoritma ini memiliki struktur akar/ *root*, simpul/*node*, dan simpul daun/ *node leaf*. Hasil klasifikasi dan prediksi yang dihasilkan dengan algoritma ini tercipta dari hubungan yang terbentuk antara variabel atribut x dengan variabel target y. Setiap struktur yang dihasilkan oleh *decision tree* berisikan hal yang berbeda-beda sehingga menciptakan struktur hierarki[13].

Pohon keputusan biasa digunakan karena sederhana dan dirancang untuk data klasifikasi. Namun sekarang ini, algoritma ini lebih sering digunakan untuk variabel kontinu[14]. Sama halnya dengan algoritma lain, *decision tree* juga memiliki kelebihan dan kekurangan[15]. Kekuranngan dari algoritma ini yaitu:

1. Data training cenderung dilebih-lebihkan, sehingga hasil yang dihasilkan dapat tidak sesuai jika diterapkan untuk keseluruhan data.
2. Rumit dalam melakukan prediksi di luar batas minimal dan maksimal pada variabel dependen dalam data training.

D. Penelitian Terdahulu

Penelitian yang berkaitan dengan prediksi penyakit jantung sebelumnya telah banyak dilakukan dan dipublikasikan. Penelitian-penelitian tersebut tidak terbatas pada algoritma *decision tree* dan *logistic regression* saja. Keberadaan penelitian terdahulu dapat menjadi kajian dan gambaran dalam penelitian ini untuk mengetahui metode dan data yang digunakan serta model yang dihasilkan.

Tania Ciu dan Raymond Sunardi Oetama[16] melakukan penelitian terkait penyakit jantung dengan dataset yang sama menggunakan algoritma *logistic regression*. Algoritma ini digunakan untuk mencari keterkaitan antara variabel dependen dengan variabel independen. Berdasarkan penelitian, diperoleh hasil bahwa algoritma tersebut memiliki tingkat akurasi sebesar 85%. Sehingga dapat dikatakan jika algoritma *logistic regression* efektif dan efisien dalam memprediksi penyakit jantung. Menggunakan empat belas variabel yang terdaftar, hasil *sensitivity* ketika dilakukan uji performa adalah sebesar 0.80. Penelitian ini juga menyatakan bahwa algoritma *logistic regression* menunjukkan bahwa faktor utama penyebab penyakit jantung adalah jenis kelamin, tekanan darah, detak jantung, dan warna pembuluh darah.

Penelitian lain dilakukan menggunakan algoritma *decision tree* oleh Pareza Alam Jusia[17]. Penelitian ini dilakukan dengan melakukan *improve classification accuracy*, yakni pemodifikasian terhadap algoritma klasifikasi yang dipaka. Adapun, modifikasi tersebut dilakukan dengan penambahan dua metode yaitu *particle swarm optimization* dan *adaboost*. Selain itu, penilaian model diukur melalui nilai akurasi dan AUC (*area under curve*). Berdasarkan studi yang dilakukan, akurasi yang diperoleh tanpa melakukan modifikasi adalah sebesar 79.26% dengan nilai AUC 0.889. Namun, setelah algoritma *decision tree* dilakukan modifikasi menggunakan metode *particle swarm optimization*, nilai akurasi dan AUC bertambah menjadi 83.59% serta 0.916. Menggunakan metode *adaboost*, akurasi yang diperoleh sebesar 79.26% dan nilai AUC sebesar 0.955. Nilai AUC yang semakin mendekati satu menandakan pemodelan semakin baik. Dilihat dari akurasi dan nilai AUC, pemodelan algoritma *decision tree* semakin baik setelah dilakukan modifikasi dengan metode *particle swarm optimization*.

III. METODOLOGI

Penelitian dengan judul "Perbandingan Algoritma *Logistic Regression* dengan *Decision Tree* untuk Memprediksi Penyakit Jantung Menggunakan Bahasa Pemrograman Pyhton" menggunakan metodologi data mining CRISP-DM. Metodologi ini terdiri dari 6 tahapan yang harus dilakukan mulai dari *business understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation*, dan *deployment*[18].

[13] Wijaya, Y.A, A. Bachtiar, Kaslani & Nining R. *Analisa Klasifikasi menggunakan Algoritma Decision Tree pada Data Log Firewall*. Vol 9 No 3 (2021): Jursima Vol. 9 No. 3, Desember Tahun 2021

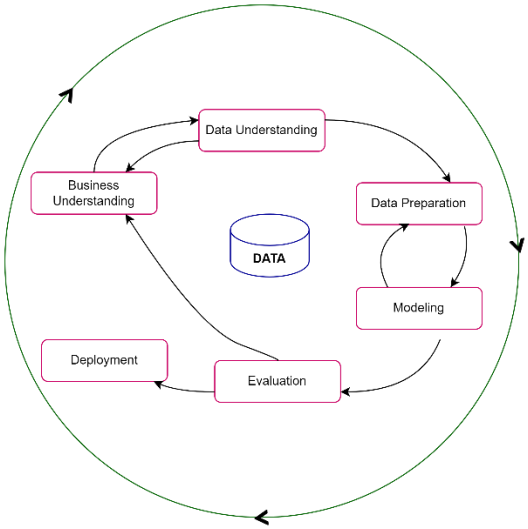
[14] Jiao, S.R. Song, J. Liu, B. A *Review of Decision Tree Classification Algorithms for Continuous Variables*. 2020

[15] T R, Prajwala. *A Comparative Study on Decision Tree and Random Forest Using R Tool*. IJARCCE. 196-199.

[16] Ciu, T., & Oetama, R. S. (2020). Logistic Regression Prediction Model for Cardiovascular Disease. *IJNMT (International Journal of New Media Technology)*, *7*(1), 33-38.

[17] P. A. Jusia, "Analisis Komparasi Pemodelan Algoritma Decision Tree Menggunakan Metode Particle Swarm Oprimization dan Metode Adaboost untuk Prediksi Awal Penyakit Jantung," *Seminar Nasional Sistem Informasi* , 2018.

[18] Hasanah, M.A, Soim, S & Handayani, A.S. Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. Journal of Applied Informatics and Computing (JAIC) Vol.5, No.2, Desember 2021, pp. 103~108



Gambar 1. Metode CRISP-DM

A. *Business Understanding*

Pada tahap ini, dilakukan pemahaman terhadap tujuan dilakukannya penelitian. Tujuan tersebut dipahami dan diperoleh dengan analisa dari sudut pandang bisnis. Kemudian, dari tujuan tersebut dilakukan pendefinisiain terkait permasalahan yang akan diselesaikan serta cara untuk menjawab permasalahan tersebut. Tahap *business understanding* sebagai langkah pertama harus dilakukan dengan matang karena berkaitan dengan pembentukan strategi untuk penyelesaian masalah.

B. *Data Understanding*

Menyelesaikan sebuah masalah untuk memperoleh tujuan tentunya memerlukan peran sebuah data sebagai pendukung untuk mencapai solusi dari permasalahan yang diangkat. Tahap *data understanding* harus dilakukan dengan teliti karena peneliti harus memahami data-data yang digunakan agar sesuai dengan permasalahan yang dihadapi. *Data understanding* dapat dilakukan dengan cara mengumpulkan, menganalisa, mendiskripsikan, mengevaluasi dan mengelompokkan data.

C. *Data Preparation*

Pada dasarnya, tahap kedua dan ketiga dalam metode CRISP-DM saling berkaitan. Setelah data dipahami dalam tahap kedua, selanjutnya data-data yang digunakan dipersiapkan untuk dioleh lebih lanjut lagi. Tindakan dalam persiapan data yang dapat dilakukan seperti: memilih variabel, mengecek dan menghapus nilai *null*, melakukan perubahan jenis data, memvisualisasikan dan mencari hubungan antar variabel.

D. *Modeling*

Setelah semua persiapan dilakukan, selanjutnya algoritma dapat langsung diaplikasikan kepada data yang bersangkutan. Algoritma yang digunakan haruslah yang sesuai dengan data, agar hasil yang diperoleh maksimal. Oleh karena itu, sebelum membuat sebuah model alangkah baiknya memahami terlebih dahulu algoritma yang akan digunakan. Adapun, hasil dari tahap *modeling* adalah pola yang terbentuk dari data yang digunakan. Pada penelitian ini,

E. *Evaluation*

Selanjutnya, model-model yang terbentuk dilakukan evaluasi. Tujuan dari tahap kelima adalah untuk mencari kesimpulan apakah model yang terbentuk menjawab permasalahan dan sesuai dengan tujuan dari tahap *business understanding*. Pada tahap ini, sebuah model juga dievaluasi dari segi kualitas serta kuantitas. Selain itu, jika algortima yang digunakan lebih dari satu maka perbandingan dilakukan pada tahap ini. Validitas sebuah model juga ditentukan melalui tahap *evaluation*, sebelum akhirnya dapat dipublikasikan.

F. *Deployment*

Sesuai dengan namanya, tahap terakhir ini berkaitan dengan penerapan model yang dihasilkan dalam kehidupan sehari-hari sesuai dengan lingkup bisnis penelitian. Contoh paling sederhana dari penerapan penerapan model yang tercipta adalah dengan membuat sebuah laporan, artikel, ataupun jurnal.

IV. HASIL DAN PEMBAHASAN

A. *Business Understanding*

Penelitian yang dilakukan berkaitan dengan data penyakit jantung. Tujuan dari penelitian ini adalah untuk melakukan prediksi potensi penyakit jantung yang dialami oleh masyarakat. Lebih lanjut, prediksi ini didasarkan pada gejala-gejala yang berkaitan dengan penyakit jantung.

B. *Data Understanding*

Penggunaan data dalam penelitian ini diperoleh dan diunduh dari situs *dataset* bernama Kaggle. Kaggle sendiri merupakan sebuah *platform* penyedia kumpulan *dataset* yang dapat digunakan untuk kebutuhan penelitian ataupun pembelajaran *machine learning*.

Dataset yang digunakan bernama "*Heart Disease Cleveland UCI*" dan diunduh melalui tautan https://www.kaggle.com/cherngs/heart-disease-cleveland-uci. *Dataset* ini telah diperbaharui dua tahun yang lalu yaitu pada 2020 dan terdiri dari 14 kolom/ variabel dengan 297 baris. Adapun, variabel-variabel yang terdapat dalam *dataset* dapat dilihat melalui tabel di bawah.

Variabel	Deskripsi
Age	Usia pasien.
Sex	Jenis kelamin (Laki-laki = 1, Perempuan = 2).
Cp	Rasa sakit pada dada. Dibagi menjadi empat: (Typical angina = 0, Atypical angina = 1, Non-anginal pain = 2, Asymptomatic 0-3 = 3).
Trestbps	Tekanan darah ketika istirahat dalam mmHg.
Chol	Kadar kolesterol dalam mg/dL.
Fbs	Gula darah setelah puasa lebih dari 120mg/dl. (Ya = 1, Tidak = 0)
Restecg	Pemeriksaan EKG ketika pasien dalam posisi berbaring. (Normal = 0, ST-T abnormal = 1, Left ventricular = 2)
Thalach	Detak jantung maksimum per menit.
Exang	Latihan yang diinduksi angina. (Ya = 1, Tidak = 0)
Oldpeak	Penurunan ST depresi ketika beristirahat, yang disebabkan oleh aktivitas olahraga.
Slope	Kemiringan segmen ST (Upsloping = 0, Flat = 1, Downsloping = 2)
Ca	Jumlah pembuluh darah yang diwarnai oleh flourosopy.
Thal	Talasemia/ kelainan darah. (Normal = 0, Fixed defect = 1, Reversable defect = 2)
Condition	Memiliki penyakit jantung. (No disease = 0, Disease = 1)

Setelah mengenali variabel-variabel yang ada, dilakukan pembuatan *dummy dataset* yang akan digunakan untuk EDA yang dilanjutkan dengan pengubahan data dalam *dummy*.

```
# dummy dataset (digunakan untuk EDA)
heart_dis = pd.read_csv("heart_cleveland_upload.csv")
dummmy_df = heart_dis # ubah nama
```

```
# mengubah data
dummmy_df['condition'] = dummmy_df['condition'].map({0: 'Normal',
                                                     1: 'Heart Attack'})
dummmy_df['sex']=dummmy_df['sex'].map({0:'Female',
                                       1:'Male'})
dummmy_df['cp']=dummmy_df['cp'].map({0:'typical angina',
                                     1:'atypical angina',
                                     2:'non-anginal pain',
                                     3:'asymptomatic'})
dummmy_df['restecg']=dummmy_df['restecg'].map({0:'Normal',1:'ST-T abnormal',
                                               2:'Left ventricular'})
dummmy_df['exang']=dummmy_df['exang'].map({0:'no',
                                           1:'yes'})
dummmy_df['thal']=dummmy_df['thal'].map({0:'Normal',
                                         1:'Fixed defect',
                                         2:'reversable defect'})
```

Gambar 2. Data understanding

C. *Data Preparation*

Tahap persiapan data dalam penelitian ini dibagi menjadi beberapa bagian lagi. Pertama, dilakukan pengecekan data yang hilang/ *missing value* menggunakan fungsi .isnull().

```
# mencari missing value
heart_disease.isnull().sum()
```

```
age          0
sex          0
cp           0
trestbps     0
chol         0
fbs          0
restecg      0
thalach      0
exang        0
oldpeak      0
slope        0
ca           0
thal         0
condition    0
dtype: int64
```

Gambar 3. Mengecek missing value

Berdasarkan pengecekan, diperoleh hasil bahwa seluruh baris dari *dataset* yang digunakan tidak ada yang bersifat *null*. Langkah selanjutnya yang dilakukan adalah mengecek duplikasi data untuk memastikan tidak ada data terduplikat. Pengecekan dilakukan menggunakan fungsi *.duplicated()* dan didapatkan hasil bahwa data terhindar dari duplikasi.

```
# mencari duplikasi data
heart_disease.duplicated().sum()
```

```
0
```

Gambar 4. Mengecek duplikasi data

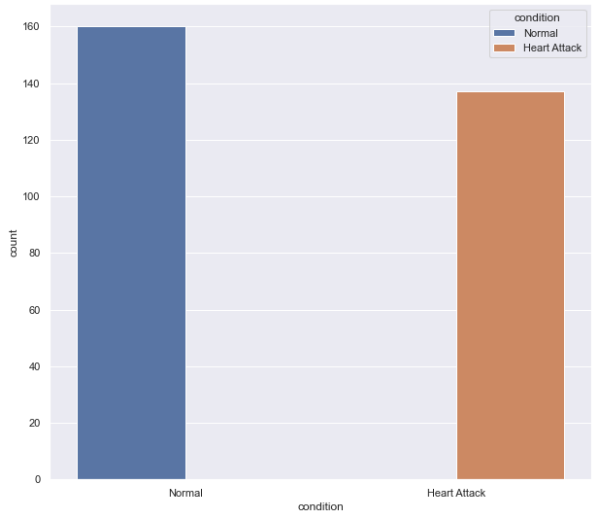
Sebagai upaya memahami data yang digunakan, dilakukan perhitungan statistik pada keempat belas variabel. Melalui perhitungan statistik, peneliti dapat mengetahui jumlah data, nilai mean, standar devisiasi. nilai minimal, serta nilai maximal dari seluruh variabel.

```
# melihat data dengan perhitungan statistik
heart_disease.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
age	297.0	54.542088	9.049736	29.0	48.0	56.0	61.0	77.0
sex	297.0	0.676768	0.468500	0.0	0.0	1.0	1.0	1.0
cp	297.0	2.158249	0.964859	0.0	2.0	2.0	3.0	3.0
trestbps	297.0	131.693603	17.762806	94.0	120.0	130.0	140.0	200.0
chol	297.0	247.350168	51.997583	126.0	211.0	243.0	276.0	564.0
fbs	297.0	0.144781	0.352474	0.0	0.0	0.0	0.0	1.0
restecg	297.0	0.996633	0.994914	0.0	0.0	1.0	2.0	2.0
thalach	297.0	149.599327	22.941562	71.0	133.0	153.0	166.0	202.0
exang	297.0	0.326599	0.469761	0.0	0.0	0.0	1.0	1.0
oldpeak	297.0	1.055556	1.166123	0.0	0.0	0.8	1.6	6.2
slope	297.0	0.602694	0.618187	0.0	0.0	1.0	1.0	2.0
ca	297.0	0.676768	0.938965	0.0	0.0	0.0	1.0	3.0
thal	297.0	0.835017	0.956690	0.0	0.0	0.0	2.0	2.0
condition	297.0	0.461279	0.499340	0.0	0.0	0.0	1.0	1.0

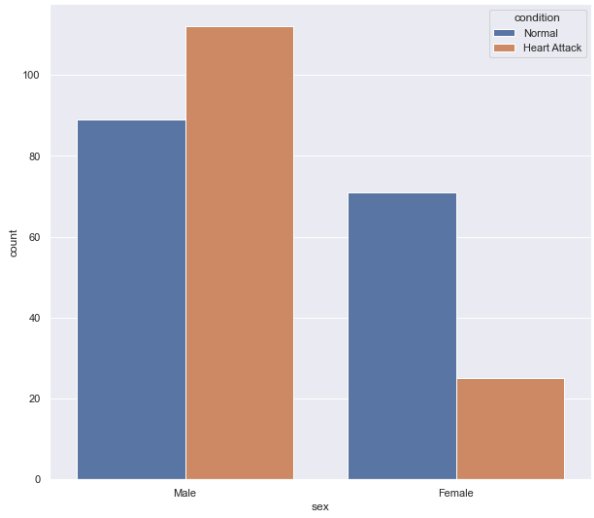
Gambar 5. Perhitungan statistik data

Tahapan persiapan data kemudian dilanjutkan dengan melakukan eksplorasi pada setiap variabel melalui visualisasi.



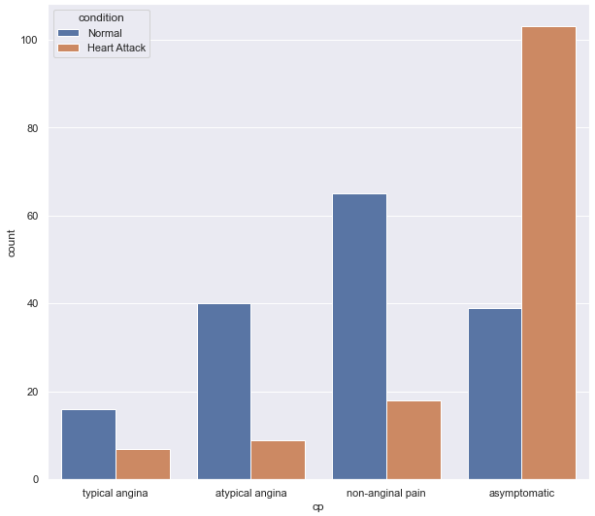
Gambar 6. Visualisasi variabel condition

Grafik di atas merupakan visualisasi dari variabel “condition”. Berdasarkan grafik, warna biru merepresentasikan kondisi normal atau pasien tanpa penyakit jantung. Sedangkan, warna oranye merepresentasikan kondisi pasien yang memiliki penyakit jantung. Dapat disimpulkan bahwa jumlah pasien normal lebih banyak dari pada pasien penyakit jantung, yaitu sebanyak 160.



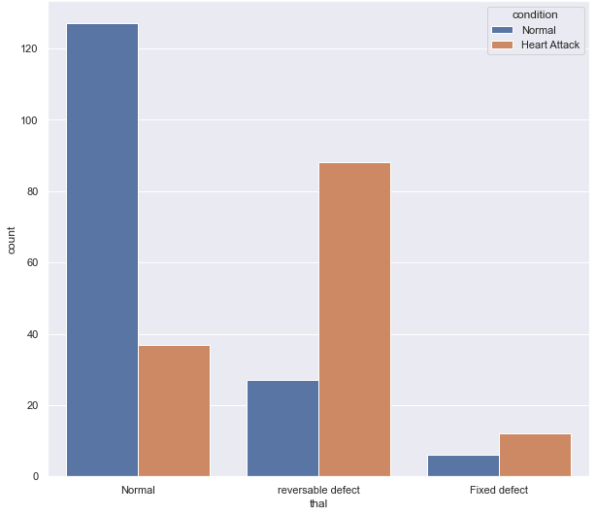
Gambar 7. Visualisasi variabel sex

Grafik di atas merupakan visualisasi dari variabel “sex” berdasarkan kondisi pasien. Dari grafik tersebut dapat dilihat bahwa pasien dengan penyakit jantung cenderung lebih banyak dialami oleh laki-laki,



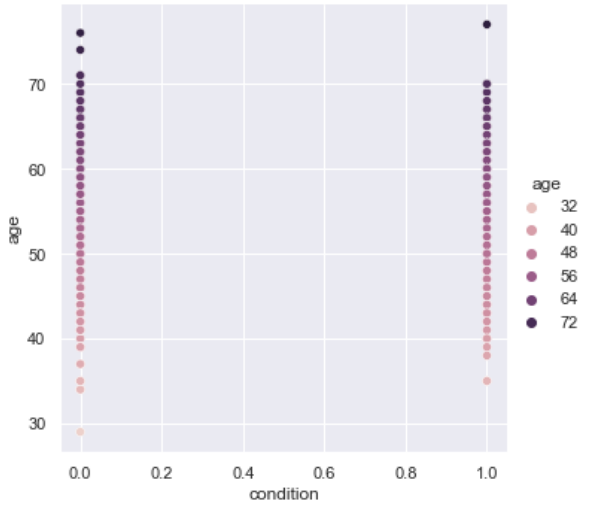
Gambar 8. Visualisasi variabel CP

Grafik di atas merupakan visualisasi dari variabel “cp” berdasarkan kondisi pasien. Dari grafik tersebut dapat dilihat bahwa pasien dengan penyakit jantung cenderung mengalami sakit dada dengan tipe Asymptomatic.



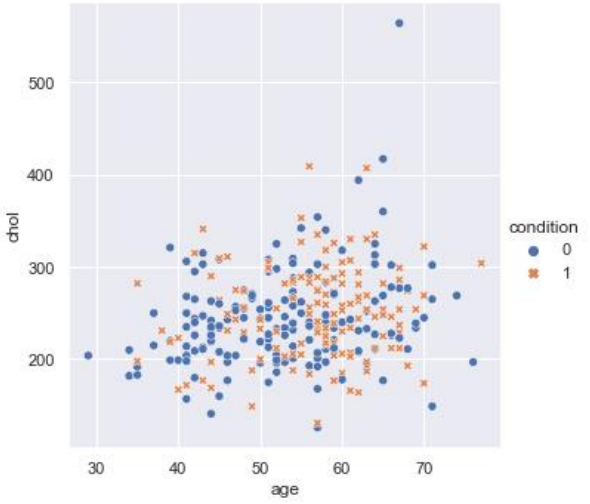
Gambar 9. Visualisasi variabel Thal

Grafik di atas merupakan visualisasi dari variabel “thal” berdasarkan kondisi pasien. Dari grafik tersebut dapat dilihat bahwa pasien dengan kondisi normal atau tidak memiliki penyakit jantung umumnya tidak memiliki talasemia (kelainan darah). Sedangkan pasien dengan penyakit jantung umumnya berada di kategori “Reversable defect”.



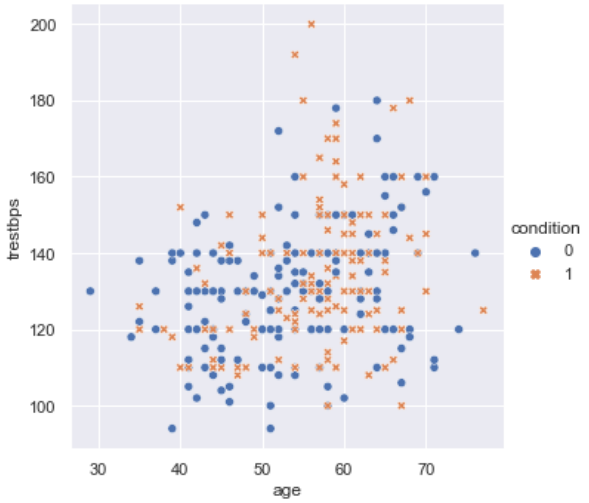
Gambar 10. Visualisasi variabel Condition dan Age

Visualisasi di atas merupakan visualisasi dari variabel “age” berdasarkan kondisi pasien. Warna yang semakin pekat menggambarkan usia yang semakin tua. Pada visualisasi di atas, dapat dilihat bahwa pasien yang memiliki penyakit jantung dimulai pada range usia 30-75 tahun.



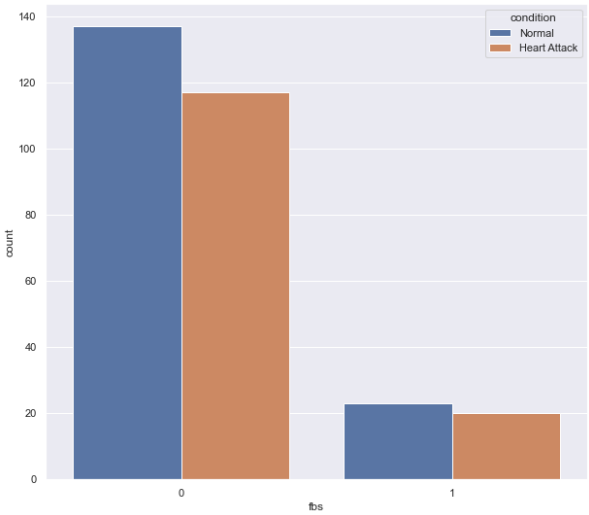
Gambar 11. Visualisasi variabel Age dan Chol

Grafik di atas merupakan visualisasi dari variabel “age” berdasarkan kondisi pasien. Dari grafik tersebut dapat dilihat bahwa pasien yang memiliki penyakit jantung ditandai dengan warna orange dan tersebar pada range usia 55-60 tahun.



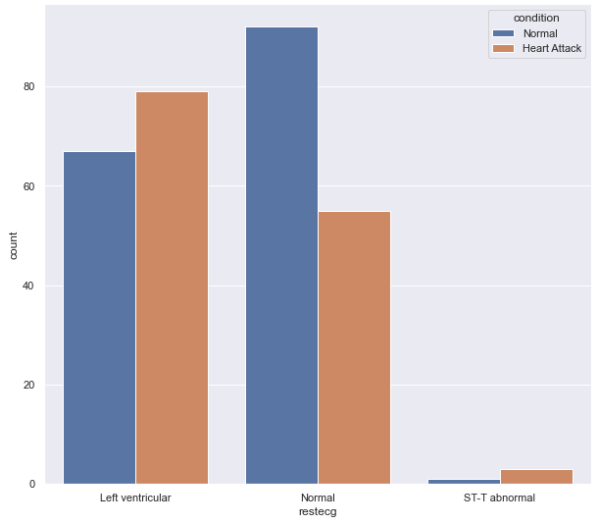
Gambar 12. Visualisasi variabel Age dan Trestbps

Grafik di atas merupakan visualisasi dari variabel “age” dan “trestbps” berdasarkan kondisi pasien. Dari grafik tersebut dapat dilihat bahwa pasien yang memiliki penyakit jantung yang ditandai dengan warna orange tersebar pada range usia 55-60 tahun dan cenderung memiliki tingkat tekanan darah saat beristirahat pada range 130-140 mmHg.



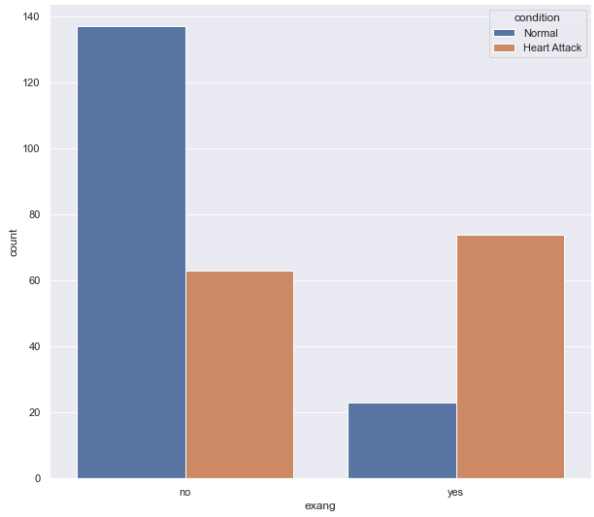
Gambar 13. Visualisasi variabel FBS

Grafik di atas merupakan visualisasi dari variabel “fbs” berdasarkan kondisi pasien. Dari grafik tersebut dapat dilihat bahwa baik pasien dengan kondisi memiliki penyakit jantung dan tidak memiliki penyakit jantung, keduanya tidak memiliki tekanan gula darah setelah puasa > 120mg/dL.



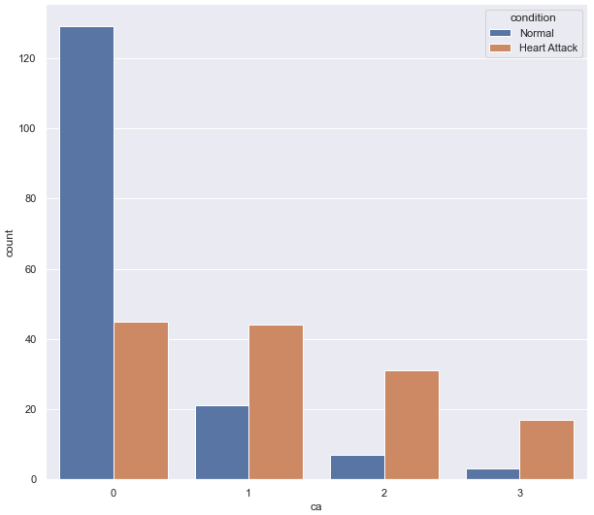
Gambar 14. Visualisasi variabel Restecg

Grafik di atas merupakan visualisasi dari variabel "restecg" berdasarkan kondisi pasien. Dari grafik tersebut dapat dilihat bahwa umumnya, pasien yang memiliki penyakit jantung ketika melakukan pemeriksaan EKG dalam posisi berbaring berada di kategori "Left ventricular". Sedangkan, pasien yang tidak memiliki penyakit jantung cenderung berada di kategori "Normal".



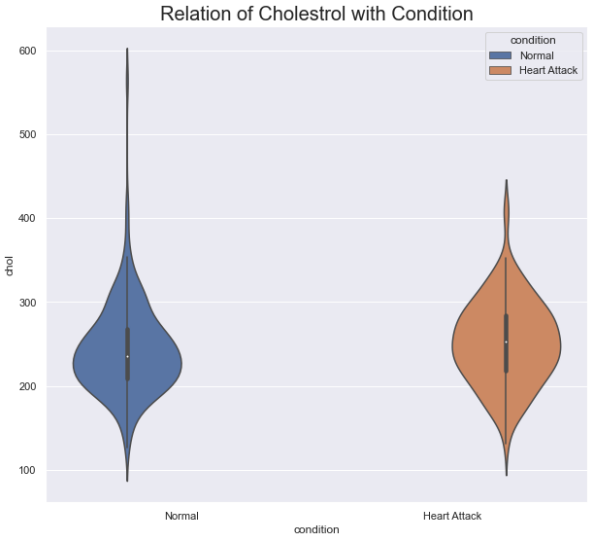
Gambar 15. Visualisasi variabel Exang

Grafik di atas merupakan visualisasi dari variabel "exang" berdasarkan kondisi pasien. Dari grafik tersebut dapat dilihat bahwa pasien yang memiliki penyakit jantung umumnya melakukan latihan yang diinduksi angina. Sedangkan pasien yang tidak memiliki penyakit jantung cenderung tidak melakukan latihan yang diinduksi angina.



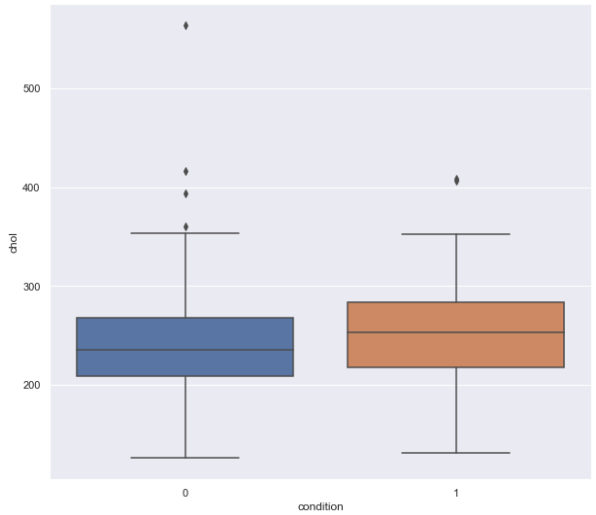
Gambar 16. Visualisasi variabel Ca

Grafik di atas merupakan visualisasi dari variabel "ca" berdasarkan kondisi pasien. Dari grafik tersebut dapat dilihat bahwa pasien dengan kondisi memiliki penyakit jantung, cenderung berada pada value 0-1 untuk jumlah pembuluh darah yang diwarnai oleh flourosopy. Sedangkan, pasien yang tidak memiliki penyakit jantung berada pada value 0.



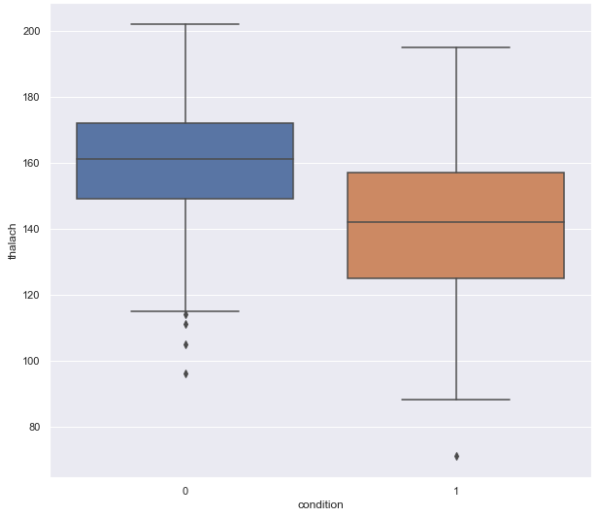
Gambar 17. Visualisasi relasi antara variabel Chol dengan Condition

Berdasarkan visualisasi data variabel "chol" berdasarkan kondisi pasien, dapat dilihat bahwa kadar kolesterol atau lemak cukup mempengaruhi kondisi pasien. Hal ini dapat dilihat pada visualisasi violin plot berwarna *orange* yang menandakan pasien dengan kondisi memiliki penyakit jantung. Semakin cembung grafik violin plot menunjukkan kepadatan data yang memiliki peluang semakin besar.



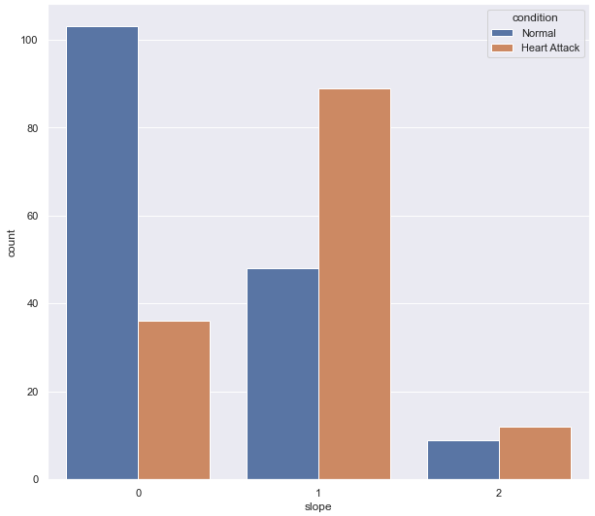
Gambar 18. Visualisasi boxplot variabel Condition dan Chol

Boxplot di atas merupakan visualisasi dari variabel "chol" berdasarkan kondisi pasien. Warna biru mengambarkan pasien dengan kondisi tidak memiliki penyakit jantung dan warna *orange* mengambarkan pasien yang memiliki penyakit jantung. Dari boxplot di atas, dapat dilihat bahwa pasien dengan kondisi memiliki penyakit jantung mempunyai nilai maximum kadar kolesterol sebesar 350 mg/dL dan nilai minimum 250 mg/dL. Selain itu, boxplot pasien penyakit jantung juga memiliki nilai outlier yang di range 400 mg/dL.



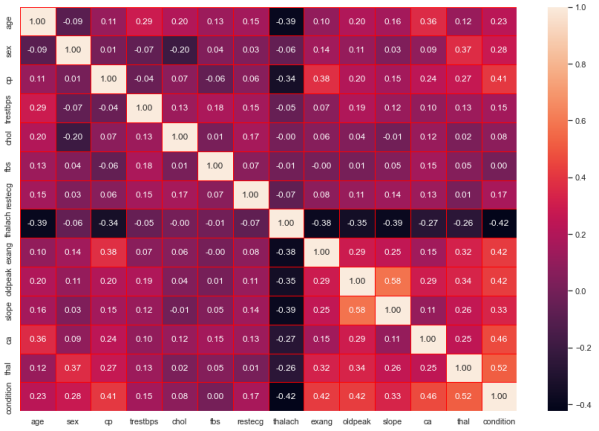
Gambar 19. Visualisasi boxplot variabel Condition dan Thalach

Boxplot di atas merupakan visualisasi dari variabel "thalach" berdasarkan kondisi pasien. Warna biru mengambarkan pasien dengan kondisi tidak memiliki penyakit jantung dan warna orange mengambarkan pasien yang memiliki penyakit jantung. Dari boxplot di atas, dapat dilihat bahwa pasien dengan kondisi memiliki penyakit jantung mempunyai nilai maximum detak jantung per menit pada range 190 dan nilai minimum 90. Selain itu, boxplot pasien penyakit jantung juga memiliki nilai outlier yang di range 70.



Gambar 20. Visualisasi variabel Slope

Grafik di atas merupakan visualisasi dari variabel "slope" berdasarkan kondisi pasien. Dari grafik tersebut dapat dilihat bahwa pasien yang memiliki penyakit jantung berada pada kategori "flat" untuk kemiringan segmen ST. Sedangkan, pasien yang tidak memiliki penyakit jantung cenderung berada pada kategori "upsloping" untuk kemiringan segmen ST.



Gambar 21. Heatmap korelasi antar variabel

Berdasarkan heatmap di atas, dapat disimpulkan bahwa *variabel cp, thal, ca, oldpeak, exang, slope, sex, age, restecg, trestbps, chol,* dan *fbs* memiliki korelasi yang positif dengan variabel *condition*. Sedangkan, *thalach* memiliki korelasi negatif dengan variabel *condition*.

```
# Memilih variabel independen (target) dan dependen
Y= heart_disease['condition'] #independen
X=heart_disease.drop(['condition','age','fbs','oldpeak'],axis=1) #dependen

X.shape
```

```
(297, 10)
```

Gambar 22. Memilih variabel independen dan dependen

Berdasarkan hasil heatmap dan visualisasi (EDA). Ditentukan bahwa, variabel *condition* merupakan variabel independen (target) dan variabel *age, fbs,* dan *oldpeak* merupakan variabel dependen.

```
# Split data
# 80% training and 20% testing.
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size = 0.2,
                                                    random_state = 0)

print("Shape of x_train :", x_train.shape)
print("Shape of x_test :", x_test.shape)
print("Shape of y_train :", y_train.shape)
print("Shape of y_test :", y_test.shape)
```

```
Shape of x_train : (237, 10)
Shape of x_test : (60, 10)
Shape of y_train : (237,)
Shape of y_test : (60,)
```

Gambar 23. Membagi data training dan testing

Pada gambar di atas, dilakukan *splitting data*. Data dibagi ke dalam perbandingan 80:20. 80% sebagai data *training set* dan 20% sebagai data *testing set.*

D. *Modelling*

A. *Logistic Regression*

```
# Print Logistic Regression
print('Accuracy of Logistic Regression on training set: {:.3f}'.format(lr.score(x_train, y_train)))
print('Accuracy of Logistic Regression on testing set: {:.3f}'.format(lr.score(x_test, y_test)))
print('\n')

# Metrics module for accuracy calculation
print("How often the classifier model correct?")
print("Accuracy: ", metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy of Logistic Regression on training set: 0.848
Accuracy of Logistic Regression on testing set: 0.850


How often the classifier model correct?
Accuracy:  0.85
```

Gambar 24. Perhitungan akurasi Logistic Regression

Hasil prediksi dengan menggunakan algoritma *Logistic Regression* menunjukan hasil akurasi sebesar 0,85. Hasil prediksi menunjukan bahwa regresi logistik cukup baik untuk digunakan dalam prediksi data penyakit jantung.

```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.72      0.95      0.82        22
           1       0.97      0.79      0.87        38

    accuracy                           0.85        60
   macro avg       0.85      0.87      0.85        60
weighted avg       0.88      0.85      0.85        60
```

Gambar 25. Classification report Logistic Regression

Berikut merupakan hasil *classification report* untuk algoritma *Logistic Regression. Accuracy* menunjukan rasio prediksi keseluruhan data yang benar (baik positif atau negative). *Precission* merupakan rasio *true positive* (TP) terhadap keseluruhan prediksi positive. Recall merupakan rasio *true positive* (TP) terhadap keseluruhan data yang benar-benar bernilai positif. *F1-score* merupakan mean dari precission dan recall. Dapat dilihat dari hasil, bahwa algoritma *Logistic Regression* memberikan hasil yang baik dalam hal *accuracy*, *precision*, *recall*, dan *f1-score* karena bernilai > 70.

B. *Decision Tree*

```
Accuracy of Decision Tree on training set: 1.000
Accuracy of Decision Tree on testing set: 0.733


How often the classifier model correct?
Accuracy:  0.7333333333333333
```

Gambar 26. Perhitungan akurasi Decision Tree

Hasil prediksi dengan menggunakan algoritma *Decision Tree* menunjukan hasil akurasi sebesar 0,73. Hasil prediksi menunjukan bahwa *decision tree* cukup baik untuk digunakan dalam prediksi data penyakit jantung.

```
print(classification_report(y_test, y_pred))
```

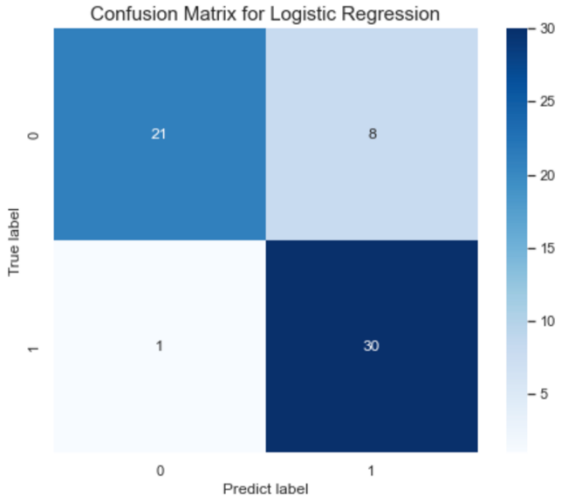
```
              precision    recall  f1-score   support

           0       0.61      0.77      0.68        22
           1       0.84      0.71      0.77        38

    accuracy                           0.73        60
   macro avg       0.73      0.74      0.73        60
weighted avg       0.76      0.73      0.74        60
```

Gambar 27. Classification report Decision Tree

Berikut merupakan hasil *classification report* untuk algoritma *Decision Tree. Accuracy* menunjukan rasio prediksi keseluruhan data yang benar (baik positif atau negative). *Precission* merupakan rasio *true positive* (TP) terhadap keseluruhan prediksi positive. Recall merupakan rasio *true positive* (TP) terhadap keseluruhan data yang benar-benar bernilai positif. *F1-score* merupakan mean dari precission dan recall. Dapat dilihat dari hasil, bahwa algoritma *Logistic Regression* memberikan hasil yang cukup baik dalam hal *accuracy*, *precision*, *recall*, dan *f1-score* karena bernilai > 50. Namun jika dibandingkan dengan performa dari algoritma *logistic regression* maka algoritma *decision tree* menghasilkan performa yang lebih buruk.

E. *Evaluation*

A. *Logistic Regression*



Gambar 28. Confusion Matrix Logistic Regression

Pada confusion matrix untuk data penyakit jantung di atas, dapat dilihat bahwa terdapat 4 hasil klasifikasi, yaitu *True Positive* (TP) = 21. *True Positive* merepresentasikan pasien yang diprediksi benar memiliki penyakit jantung. Selanjutnya, terdapat *False Postive* (FP) = 8, *False Negative (FN)* = 1, dan *True Negative* (TN) = 30.

Cross Validation

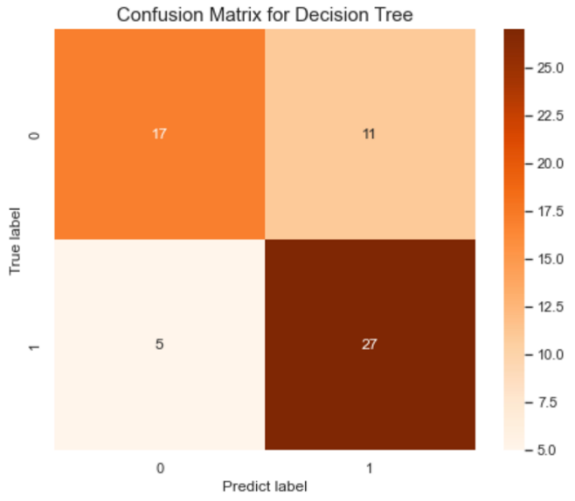
```
CROSS_final = cross_val_score(lr, x_train, y_train, cv=10).mean()
CROSS_final
```

```
0.8342391304347826
```

Gambar 29. Cross Validation Logistic Regression

Pada gambar di atas, dilakukan *cross validation* untuk menguji dan mengevalusasi kembali model Logistic Regression. Dari *cross validation* ini, di dapatkan akurasi untuk model Logistic Regression sebesar 0.83423. Dari hasil ini, dapat disimpulkan bahwa nilai akurasi pada algoritma Logistic Reggresion lebih rendah ketika diuji dengan menggunakan metode *cross validation*.

B. *Decision Tree*



Gambar 30. Confusion Matrix Decision Tree

Pada confusion matrix untuk data penyakit jantung di atas, dapat dilihat bahwa terdapat 4 hasil klasifikasi, yaitu *True Positive* (TP) = 17. *True Positive* merepresentasikan pasien yang diprediksi benar memiliki penyakit jantung. Selanjutnya, terdapat *False Postive* (FP) = 11, *False Negative (FN)* = 5, dan *True Negative* (TN) = 27.

Cross Validation

```
CROSS_final = cross_val_score(dt, x_train, y_train, cv=10).mean()
CROSS_final
```

```
0.7204710144927535
```

Gambar 31. Cross Validation Decision Tree

Pada gambar di atas, dilakukan *cross validation* untuk menguji dan mengevalusasi kembali model Decision Tree. Dari *cross validation* ini, di dapatkan akurasi untuk model Decision Tree sebesar 0.72047. Dari hasil ini, dapat disimpulkan bahwa nilai akurasi pada algoritma Decision Tree lebih rendah ketika diuji dengan menggunakan metode *cross validation*.

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	84.810127	85.000000
1	Decision Tree	100.000000	73.333333

Gambar 32. Comparasion 2 Algoritma

Berdasarkan gambar di atas, dapat kita lihat bahwa Logstic Regression menjadi algoritma dengan performa yang lebih baik digunakan untuk memprediksi penyakit jantung pada dataset Heart Disease Cleveland UCI. Hal ini dilihat dari nilai akurasi testing yang lebih tinggi, yaitu 85%. Selain itu, nilai *True Positive* (TP) pada algoritma Logistic Regression lebih banyak yaitu sebesar 21 dibandingkan dengan algoritma Decision Tree.

F. *Deployment*

Berdasarkan penelitian, dapat disimpulkan bahwa untuk memprediksi penyakit jantung pasien pada dataset "Heart Disease Cleveland UCI" dapat dilakukan dengan menggunakan algoritma Logistic Regression. Variabel yang diduga cukup akurat untuk memprediksi penyakit jantung adalah variabel *variabel cp, thal, ca, oldpeak, exang, slope, sex, age, restecg, trestbps, chol,* dan *fbs.* Hal ini dikarenakan variabel-variebel tersebut memiliki korelasi yang positif dengan variabel *condition*. Hasil penelitian dapat bermanfaat dalam dunia kesehatan untuk membantu memprediksi faktor apa saja yang menyebabkan penyakit jantung.

V. KESIMPULAN

A. *Kesimpulan*

Hasil prediksi penyakit jantung dengan membandingkan Algoritma Logistic Regression dan Algoritma Decision Tree pada penelitian menunjukkan nilai akurasi tertinggi sebesar 85%. Peneliti menyimpulkan bahwa saat membuat prediksi pada dataset penyakit jantung, algoritma *Logistic Regression* mengungguli algoritma *Decision Tree*, dilihat dari hasil nilai akurasi testing yang lebih tinggi, yaitu 85%. Selain itu, nilai *True Positive* (TP) pada algoritma Logistic Regression lebih banyak yaitu sebesar 21 dibandingkan dengan algoritma Decision Tree.

Oleh karena itu, dapat disimpulkan bahwa model prediksi *Logistic Regression* merupakan algoritma yang

lebih baik untuk dapat digunakan dalam memprediksi dini penyakit jantung berdasarkan gejala.

B. Saran

Pada penelitian ini, peneliti melakukan perbandingan menggunakan 2 algoritma yaitu Decision Tree dan Logistic Regression sehingga pada penelitian selanjutnya dapat dikembangkan dengan menambahkan beberapa algoritma klasifikasi maupun clustering lainnya agar peneliti selanjutnya mendapatkan perbandingan yang lebih beragam dan perlu dilakukan pengujian dengan menggunakan data dalam jumlah yang lebih besar agar perbandingan akurasi dan model lebih baik lagi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Bapak Rudi Sutomo, selaku dosen *Data Modelling* kelas C Sistem Informasi Universitas Multimedia Nusantara, yang telah meluangkan waktunya untuk mengajar, memberikan nasehat, arahan, dan memberikan ilmunya selama menempuh studi analisis di semester ini.

DAFTAR PUSTAKA

[1] WHO. Cardiovascular diseases (CVDs). June 11, 2021. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [Accessed 9 October 2021].

[2] Shoufiah, R. Hubungan Faktor Resiko dan Karakteristik Penderita dengan Kejadian Penyakit Jantung Koroner. Mahakam Nursing Journal Vol 1, No. 1, May 2016 : 17-26

[3] R. E. H. Patriyani and D. F. Purwanto, "Faktor Dominan Risiko Terjadinya Penyakit Jantung (PJK)," *Keperawatan Global*, vol. 1, no. 1, p. 24, 2016.

[4] Halodoc, R. (2021, November 25). Detak jantung normal seseorang ditentukan berdasarkan usia hingga aktivitas yang dilakukan. Simak selengkapnya. halodoc. https://www.halodoc.com/artikel/berapa-detak-jantung-normal-berdasarkan-usia

[5] I. Ramadini and S. Lestari, "Hubungan Aktivitas Fisik dan Stress dengan Nyeri Dada Pasien Penyakit Jantung Koroner," *Human Care*, vol. 2, no. 3, 2017.

[6] -. (2011, September 6). Kematian Akibat Serangan Jantung Banyak Menimpa Atlet Muda. detikHealth. https://health.detik.com/berita-detikhealth/d-1716590/kematian-akibat-serangan-jantung-banyak-menimpa-atlet-muda

[7] Escamila, A.K. Hassani, A. H & Andres, E. Classification models for heart disease prediction using feature selection and PCA. Volume 19, 2020.

[8] Telaumbanua, F. D., Hulu, P., Nadeak, T. Z., Lumbantong, R. R., & Dharma, A.. Penggunaan Machine Learning Di Bidang Kesehatan. JURNAL TEKNOLOGI DAN ILMU KOMPUTER PRIMA (JUTIKOMP), 2(2), 391-399. 2020. https://doi.org/10.34012/jutikomp.v2i2.657

[9] Nurdiansah, S.N & Khikmah, Laelatul. *Binary Logistic Regression Analysis of Variables that Influence Poverty in Central Java.* Vol. 1, No. 1, March 2020

[10] E. Y. Boateng, and D. A. Abaye, *"A Review of the Logistic Regression Model with Emphasis on Medical Research"*. Journal of Data Analysis and Information Processing, Vol.7. No.4, pp.190-207, 2019.

[11] Edgar, T.W. & Manz, D.O. *Research Methods for Cyber Security*. 2017

[12] Rianto, H. *Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software.* Journal of Software Engineering, Vol. 1, No. 1, April 2015

[13] Wijaya, Y.A, A. Bachtiar, Kaslani & Nining R. *Analisa Klasifikasi menggunakan Algoritma Decision Tree pada Data Log Firewall.* Vol 9 No 3 (2021): Jursima Vol. 9 No. 3, Desember Tahun 2021

[14] Jiao, S.R. Song, J. Liu, B. *A Review of Decision Tree Classification Algorithms for Continuous Variables*. 2020

[15] T R, Prajwala. *A Comparative Study on Decision Tree and Random Forest Using R Tool*. IJARCCE. 196-199.

[16] Ciu, T., & Oetama, R. S. (2020). Logistic Regression Prediction Model for Cardiovascular Disease. *IJNMT (International Journal of New Media Technology)*, *7*(1), 33-38.

[17] P. A. Jusia, "Analisis Komparasi Pemodelan Algoritma Decision Tree Menggunakan Metode Particle Swarm Oprimization dan Metode Adaboost untuk Prediksi Awal Penyakit Jantung," *Seminar Nasional Sistem Informasi* , 2018.

[18] Hasanah, M.A, Soim, S & Handayani, A.S. *Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir*. Journal of Applied Informatics and Computing (JAIC) Vol.5, No.2, Desember 2021, pp. 103~108