ANALISIS DAN PENGEMBANGAN SISTEM DETEKSI PENYAKIT JANTUNG DENGAN MACHINE LEARNING

Alfian Imran
Department of Computer engineering
Telkom University
Bandung, Indonesia
alfianimran@student.telkomuniversity.ac.id

Abstrak - Penyakit Jantung Koroner adalah gangguan fungsi jantung akibat otot jantung kekurangan darah karena penyumbatan atau penyempitan pada pembuluh darah koroner akibat kerusakan lapisan dinding pembuluh darah (Aterosklerosis). Hasil Riset Kesehatan Dasar (Riskesdas) tahun 2013 menunjukan bahwa prevalensi tertinggi untuk penyakit kardiovaskuler di Indonesia adalah Penyakit Jantung Koroner sebesar 2.650.340 orang (0,5%). Hasil Riskesdas tahun 2018 menunjukkan bahwa prevalensi untuk penyakit kardiovaskuler di Indonesia meningkat menjadi 1,5%. Metode yang digunakan pada Tugas Besar ini adalah K-Nearest Neighbor, Logistic Regression, dan Decision Tree. Algoritma terbaik untuk kasus ini adalah KNN dan Logistic Regression dengan akurasi 0,8833 sementara akurasi Decision Tree sendiri adalah 0,6833.

I. PENDAHULUAN

Manusia pasti menginginkan untuk memiliki hidup yang sehat dan umur yang panjang untuk menjalani hidup. Namun, pada kenyataannya manusia dihadapkan pada risiko terkena penyakit. Mulai dari penyakit kronis hingga penyakit mematikan yang akan membahayakan hidupnya. Salah satu dari penyakit mematikan tersebut adalah penyakit jantung koroner.

Penyakit jantung koroner adalah penyakit jantung yang disebabkan oleh penyempitan pembuluh darah arteri koroner. Penyempitan pembuluh darah arteri dapat disebabkan oleh ketidakseimbangan antara kebutuhan dan pasokan oksigen yang masuk ke dalam pembuluh darah.

Penyakit jantung koroner merupakan salah satu penyebab kematian paling tinggi di dunia. Data dari *European Cardiovascular Disease Statistics* menunjukkan 45 % kematian di Eropa disebabkan oleh penyakit jantung atau sekitar 3,9 juta orang. Penyakit jantung koroner juga merupakan salah satu penyebab kematian tertinggi pada wanita.

Di Indonesia penyakit jantung koroner merupakan penyebab kematian no 2 pada tahun 2012 yang menyebabkan 9% kematian (138.400) orang. Data Institute for Health Metrics and Evaluation menunjukkan sebanyak 14,4 % penyakit jantung koroner menyumbangkan 14,4 % penyebab kematian di Indonesia atau sekitar 2.784.064 orang. Diperkirakan pada

tahun 2024 penderita penyakit jantung koroner akan mencapai angka 6 juta orang.

Berbagai penelitian telah banyak dilakukan pada berbagai bidang untuk menekan angka kematian akibat penyakit jantung koroner. Baik dari bidang Kesehatan maupun bidang teknologi informasi. Penelitian-penelitian tersebut kebanyakan berfokus untuk dapat memprediksi seseorang menderita penyakit jantung sehingga memungkinan untuk dapat ditangani lebih awal.

Beberapa contoh penelitian yang dilakukan untuk memprediksi penyakit jantung sedini mungkin adalah dengan memanfaatkan teknologi *Artificial Intelligence (AI).* Algoritma yang digunakan pun bermacam-macam, salah satu contoh algoritma yang paling sering digunakan adalah *deep learning.* Dengan menggunakan algoritma *deep learning* maka dimungkinkan untuk memprediksi seseorang terkena penyakit jantung berdasarkan data-data yang sudah ada.

II. Landasan Teori

A. K-Nearest Neighbor

K-Nearest Neighbor adalah salah satu algoritma yang mengimplementasikan *supervised learning* dalam mengklasifikasikan data. Algoritma ini biasanya digunakan untuk mengklasifikasikan beberapa objek pada suatu dataset berdasarkan kedekatan beberapa objek dengan *k.*

Dekat atau tidaknya jarak tersebut akan dihitung oleh sistem menggunakan metode *cosine similarity [].* Berikut contoh rumus yang digunakan untuk menghitung *cosine similarity.*

$$\cos(\theta_{QD}) = \frac{\sum_{i=1}^{n} Q_i D_i}{\sqrt{\sum_{i=1}^{n}(Q_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(D_i)^2}}$$

Ket.
$\cos(\theta_{QD})$ = kemiripan Q terhadap dokumen D
Q = Data Uji
D = Data latih
n = banyak data

B. Logistic Regression

Logistic Regression adalah salah satu algoritma klasifikasi yang dapat digunakan untuk mencari hubungan antara firut (input) diskrit/kontinu dengan probabilitas hasil output diskrit tertentu. Logistic regression juga dapat dikategorikan sebagai *supervised learning.*

Logistic Regression menggunakan fungsi *logit functions* yang akan menghitung probabilitas hubungan antara *independent variables* dan *dependent variables.* Hasil tersebut kemudian akan diubah menjadi *binary values* dengan output “1” atau “0”.

C. Decision Tree

Banyak ahli mengkategorikan Decision tree sebagai *non parametric models* yang dapat digunakan untuk klasifikasi maupun regresi. Artinya,

Decision Tree dapat digunakan untuk machine learning dengan keluaran *categorical* maupun *numerical*.

Decision Tree biasanya meniru cara berpikir manusia untuk memahami data. Algoritma *Decision Tree* biasanya sangat mudah dipahami oleh orang awam karena mengambil keputusan dengan cara yang sangat simple berdasarkan kondisi tertentu.

III. Metode

1. Import Dataset

Langkah awal yang akan dilakukan adalah mengimport dataset agar dapat diproses di dalam *Google Colab*. Berikut perintah yang diinput untuk mengimport dataset.

```
data = pd.read_csv("heart_cleveland_upload.csv")
data.head(10)
```

Gambar 3.1 Import Dataset

2. Data Description

Dataset yang digunakan dalam Tugas Besar ini adalah dataset yang disediakan secara gratis oleh Cleveland UCI Machine Learning Repository dan dapat dilihat pada link berikut.
https://archive.ics.uci.edu/ml/datasets/Heart+Disease

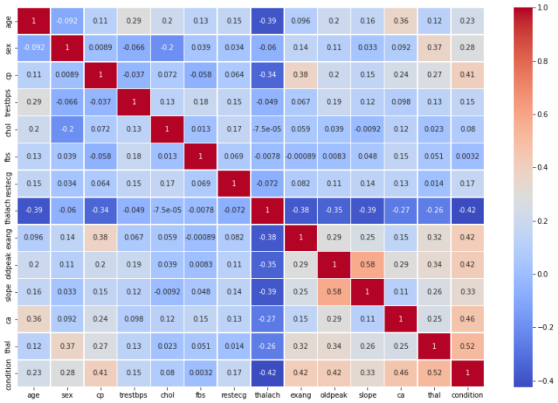
Dari 76 atribut yang tersedia, penulis hanya menggunakan beberapa atribut yang dianggap penting. Berikut atribut tersebut.

- age
- sex (1 = male; 0 = female)
- cp (Chest Pain type)
    - Value 0: typical angina
    - Value 1: atypical angina
    - Value 2: non-anginal pain
    - Value 3: asymptomatic
- trestbps (Resting Blood Pressure)
- chol (Cholestoral)
- fbs (Fasting Blood sugar)
- restecg (resting electocardiographic result.
    - Value 0: normal
    - Value 1: having ST-T Wave abnormality
    - Value 2: showing probable or definite left ventricular hypertrophy
- thalach (maximum heart rate achieve)
- exang (exercise induce angina 1=yes; 0=no)
- oldpeak (ST depression induced by excersice relative to rest)
- slope (the slope of the peak excersie ST segment)
    - Value 0: upsloping
    - Value 1: flat
    - Value 2: downsloping
- ca (number of major vessels)
- thal (0 = normal, 1 = fixed defect, 2 = reversable defect and label
- condition (0 = no disease, 1 = disease)

Dataset tersebut terdiri dari 297 data dengan 14 atribut di atas dengan 13 *independent variable* dan 1 *dependent variable*. Sementara jumlah pasien penderita penyakit jantung adalah 137 orang.

3. Data Analysis

a) Multivariate Analysis



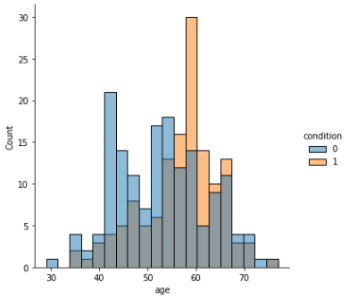
Gambar 3.2 Correlation Heatmap

Dari grafik *Correlation Heatmap* yang ditampilkan di atas dapat disimpulkan bahwa tidak ada variabel yang berkorelasi tinggi dengan atribut "*condition"*. Atribut yang dapat dianggap memiliki korelasi yang cukup signifikan adalah "thal", "ca", "slope", "oldpeak", exang", "cp", "sex", dan "age". Namun, ada satu variabel yang memiliki korelasi negative dengan *dependent variable*, yaitu atribut "thalach".

b) Univariate Analysis dan Bivariate Analysis

Univariate analysis adalah analisis yang dilakukan untuk menganalisa suatu variabel untuk mengetahui fenomena data pada variabel tersebut.

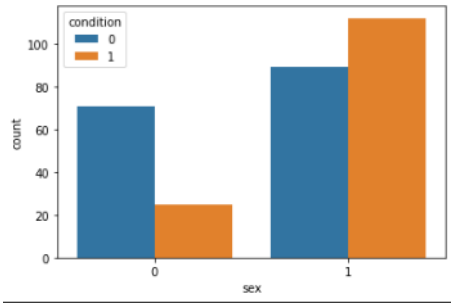
- age



Gambar 3.3 thal

Jumlah pasien dengan faktor umur sangat berpengaruh terhadap potensi penyakit jantung. Dari grafik di atas dapat dilihat pasien dengan usia antara 55-60 tahun adalah pasien penderita penyakit jantung terbanyak sementara di atas 67 dan di bawah 55 tahun memiliki kemungkinan terkecil.

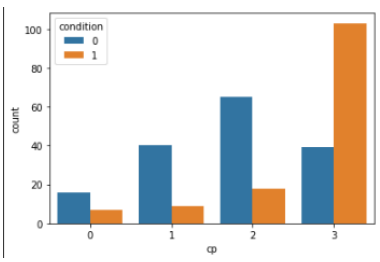
- sex



Gambar 3.4 sex

Dari yang dapat dilihat pada gambar 3.4 pasien pria adalah penderita penyakit jantung terbanyak. Namun, hal ini terjadi karena adanya ketidakseimbangan data pada dataset tersebut dengan jumlah pasien pria 201 sementara pasien wanita hanya berjumlah 96 sehingga dapat ditarik kesimpulan bahwa sex juga tidak berpengaruh besar terhadap kondisi penyakit jantung.

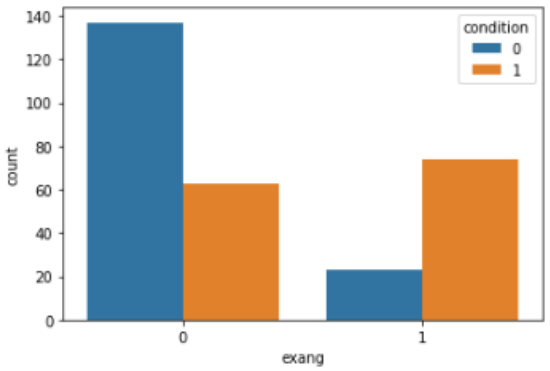
- cp



Gambar 3.5 cp

Gambar 3.5 di atas menunjukkan pasien penderita penyakit jantung terbanyak adalah pasien dengan kondisi cp type 3 atau dengan kondisi asymptomatic diikuti kondisi 2, 1, dan 0.

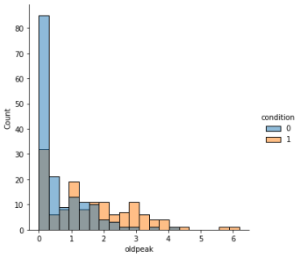
- exang



Gambar 3.6 exang

Berdasarkan penelitian-penelitian yang telah dilakukan oleh para ahli dan data yang ditunjukkan pada dataset pasien dengan kondisi *exercise induce angina* adalah pasien yang memiliki potensi besar untuk menderita penyakit jantung. Walaupun dataset menunjukkan ketidakseimbangan data antara pasien dengan *exercise induce angina* dan tidak, data tetap menunjukkan pasien penderita penyakit jantung dengan kondisi *exercise induce angina* lebih banyak.

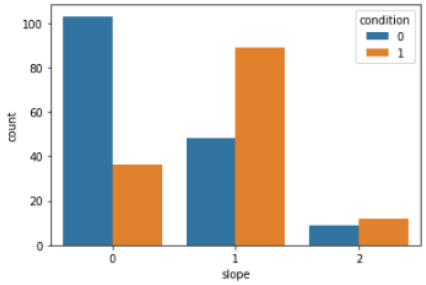
- oldpeak



Gambar 3.7 oldpeak

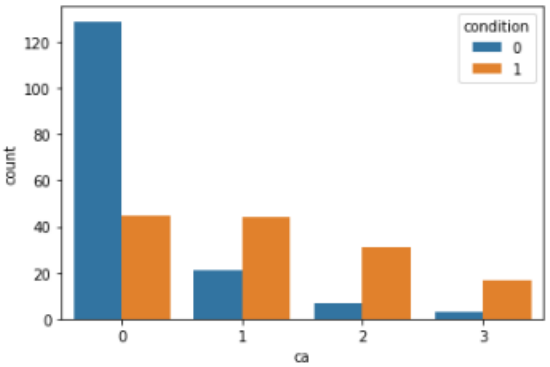
Gambar 3.7 menunjukkan distribusi pasien berdasarkan *oldpeak*. Grafik tersebut menunjukkan pasien penderita dimulai dari oldpeak = 1.

- slope



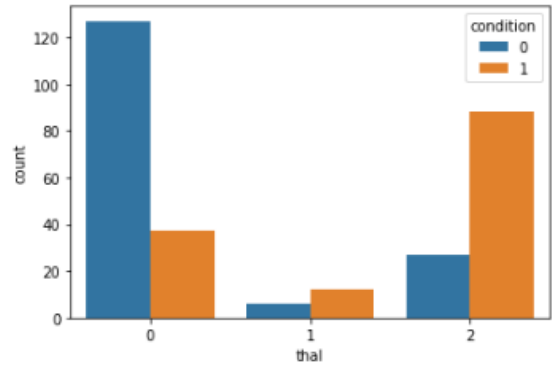
Gambar 3.8 slope

- ca



Gambar 3.9 ca

- thal



Gambar 3.10 thal

4. Data Preprocessing

a) Standard Scaler

Standard scaler adalah salah satu kelas dari Scikit Learn yang dapat digunakan untuk menormalisasi data agar penyimpangan pada data tidak lagi besar. Berikut perintah yang diberikan untuk melakukan StandardScaler.

```
x_train, x_test, y_train, y_test = train_test_split(X_std, y, test_size=0.2, random_state=10)
```

b) Train Test Split

Kemudian data dibagi menjadi dua bagian, yaitu data train dan data test. Data train adalah data yang digunakan untuk melatih model sedangkan data test adalah data yang digunakan untuk menguji performa model. Dalam Tugas Besar ini data dibagai menjadi 80% data train, dan 20% data test.

5. KNN

Training model dengan algoritma KNN mendapatkan hasil akurasi terbesar pada k=10. Akurasi yang didapat adalah 0,8833

6. Logistic Regression

Training model dengan algoritma KNN mendapatkan akurasi yang sama dengan KNN yaitu sebesar 0,8833.

7. Decision Tree

Training dengan algoritma Decision Tree mendapatkan akurasi yang terkecil yaitu 0,6833.

IV. Kesimpulan

Di antara ketiga algoritma yang digunakan, algoritma terbaik untuk kasus deteksi penyakit jantung adalah KNN dan Logistic Regression dengan nilai akurasi yang sama yaitu 0,8833.

Referensi

[1] Armin Yazdani, et al, "*A Novel Approach for Heart Disease Prediction Using Strength Scores with Significant Predictors*," BMC Medical Informatics and Decision Makinf, vol. 21, no. 194, 2021, doi: 10.1186/s12911-021-01527-5

[2] Harshit Jindal, et al., "*Heart Disease Prediction Using Machine Learning Algorithm*," IOP Conference Series : Material Science and Engineering, 2021, doi: 10.1088/1757-899X/1022/1/012072

[3] Jian Yang, Jinhan Guang, "*Prediction of Heart Disease Using Machine Learning Algorithm*," International Journal of Advanced Engineering, Management and Science, vol. 2, issue 6, June 2016

[4] Anderies, et al. "*Prediction of Heart Disease Dataset Using Machine Learning Algorithm*," Jurnal EMACS, Sept 2022, doi: 10.21512/emacsjournal.v4i3.8683