Jurnal Elektronik Ilmu Komputer Udayana
Volume 11, No 4. Mei 2023

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Analisis Algoritma Random Forest Dalam Memprediksi Penyakit Jantung Koroner

Stephania Getrudis Inaconta Sadipun**,** I Gusti Ngurah Anom Cahyadi Putra,S.T.,M.Cs

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Udayana
Jalan Raya Kampus Unud, Badung
sadipunnia@gmail.com
anom.cp@unud.ac.id

Abstract

Nowadays, there are many problems with diseases especially for internal diseases. One of the internal diseases that can affect human vital organs is coronary heart disease. Coronary heart disease is caused by excessive accumulation of fatty substances in the lining of the pulse wall of coronary vessels, and this over time is followed by various processes such as the accumulation of connective tissue, capsulation, blood clotting, which will clog blood vessels and result in the heart lacking blood. Of course, this is very dangerous for health and needs to be considered. Therefore, this study aims to test and analyze the Random Forest algorithm in accurately predicting coronary heart disease, so that with this, it is hoped that the system created can be said to be a good system so that it can provide fast and accurate treatment for those infected by this disease. The results of this study are in the form of an evaluation of the system that states that the accuracy of the system reaches 91%, then training score and cross validation which has an upward curve which means that the system learns from the data well and also the confusion matrix value which provides a high calculation of prediction correctness compared to the calculation of prediction errors, which are 24: 4 All these results are said to be good and the system is ready to be used to detect coronary heart disease.

Keywords: *Heart Disease, Detect, Predict, Random Forest, Accuracy*

1. Pendahuluan

Jantung merupakan salah satu bagian dari organ vital manusia yang perlu dijaga untuk keberlangsungan hidup manusia. Salah satu hal dalam dunia kesehatan yang saat ini menjadi sorotan adalah berbagai penyakit yang dapat dialami oleh organ vital manusia termasuk jantung. Jantung juga bisa terjangkit penyakit yang cukup mematikan dan membahayakan kesehatan salah satunya adalah penyakit jantung koroner. Penyakit jantung koroner adalah penyakit yang disebabkan oleh penumpukan zat lemak secara berlebihan di lapisan dinding nadi pembuluh koroner, dan hal ini lama kelamaan diikuti oleh berbagai proses seperti penimbunan jaringan ikat, perkapuran, pembekuan darah, dan lainya, yang mana semuanya ini akan mempersempit atau menyumbat pembuluh darah. Hal ini akan mengakibatkan otot jantung di daerah tersebut mengalami kekurangan aliran darah dan dapat menimbulkan berbagai akibat yang cukup serius bahkan bisa mengancam nyawa [1]. Hal ini perlu menjadi perhatian karena berdasarkan data dari World Health Organization (WHO) tahun 2005, dari 58 juta kematian di dunia, 17,5 juta (30%) diantaranya disebabkan oleh penyakit jantung dan pembuluh darah, terutama oleh serangan jantung (7,6 juta) dan stroke (5,7 juta). Pada tahun 2015, diperkirakan kematian penyakit jantung dan pembuluh darah di dunia meningkat menjadi 20 juta [2].

Di Indonesia, salah satu masalah kesehatan masyarakat yang sedang kita hadapi saat ini dalam pembangunan kesehatan adalah beban ganda penyakit, yaitu disatu pihak masih banyaknya penyakit infeksi yang harus ditangani, dilain pihak semakin meningkatnya penyakit tidak menular terutama penyakit jantung dan pembuluh darah. Angka kematian penyakit tidak menular meningkat dari 41.7% pada tahun 1995 menjadi 59,5% pada tahun 2007 [3].

Berdasarkan data diatas, tentu saja penyakit jantung koroner adalah penyakit yang sangat mematikan yang perlu untuk ditindak lanjuti. Salah satu cara yang dapat dilakukan adalah dengan mengetahui

keberadaan penyakit jantung koroner sedini mungkin, dengan mengembangkan sistem berbasis Machine Learning yang dapat mendeteksi keberadaan penyakit jantung pada tubuh seseorang dengan menggunakan sistem yang memiliki akurasi yang baik pula. Oleh karena itu, dalam pengembangan sistem perlu menggunakan algoritma yang bisa memprediksi penyakit jantung coroner dengan akurasi yang cukup tinggi. Salah satu algoritma yang dapat memprediksi dengan baik adalah Random Forest. Maka dari pada itu, penelitian ini akan menganalisis seberapa baik akurasi algoritma random forest dalam melakukan prediksi penyakit jantung koroner.

Penelitia ini akan menggunakan Machine Learning untuk membuat sistem yang cerdas dengan mempelajari data. Dataset yang digunakan pada sistem ini didapatkan dari website Kaggle.com. Data pada dataset ini terdiri atas data-data informasi kesehatan yang berkaitan dengan jantung dan juga berisi data keberadaan penyakit jantung koroner yang dilambangkan dengan 1 jika terdeteksi dan 0 jika tidak terdeteksi. Dataset kemudian dibagi menjadi data latih dan data tes dan dilatih dengan menggunakan algoritma Random Forest. Algoritma Random Forest itu sendiri didesain oleh J. Ross Quinlan, dinamakan Random Forest karena merupakan keturunan dari pendekatan ID3 untuk membangun pohon keputusan. Random Forest merupakan algoritma yang cocok digunakan untuk masalah klasifikasi pada machine learning dan data mining [4]. Random Forest memetakan atribut dari kelas sehingga dapat digunakan untuk menemukan prediksi terhadap data yang belum muncul. Pohon keputusan sendiri merupakan pendekatan “divide and conquer” dalam mempelajari masalah dari sekumpulan data independen yang digambarkan dalam bagan pohon [5]. Algoritma Random Forest terbentuk atas beberapa *base learner* Decision Tree, yang mana setiap hasil dari Decision Tree akan digabungkan kemudian dicari estimasi gabungannya menggunaan *Majority Voting* untuk mendapatkan hasil prediksinya.

2. Metode Penelitian

2.1. Data Acquisition (Pengumpulan Data)

Mula-mula, yang perlu dilakukan ialah mengumpulkan data-data yang akan digunakan untuk membuat sistem. Data yang digunakan adalah data Bernama “heart.csv” yang diambil dari website Kaggle.com.

2.2. Preprocessing Data

Setelah melakukan pengumpulan data, selanjutnya data perlu dilakukan preprocessing sehingga data dapat dipersiapkan dengan baik sebelum digunakan untuk membangun sistem, yang mana hal ini berguna karena panenliti hanya akan mengambil data yang memang diperlukan. Hal ini sangat penting, karena jika data yang tidak diperlukan juga ikut terambil, maka dapat memungkinkan terjadinya interpretasi hasil yang salah. Adapun beberapa preprocessing data yaitu :

- Melakukan One-Hot Endcoding (mengubah variabel kategorik dengan merepresentasikannya dalam bentuk binary dengan angka 0 dan 1)
- Menghapus variabel yang tidak diperlukan / feature selection
- Menghapus missing value
- Melakukan pemisahan feature
- Menormalisasi data
- Melakukan pembagian jumlah dataset kedalam dataset training dan dataset testing

2.3. Modelling (Pemodelan/Pembangunan Sistem)

Setelah 2 tahapan sebelumnya dilakukan, maka selanjutnya dilakukan pemodelan sistem yaitu implementasi algoritma Random Forest pada dataset. Algoritma Random Forest akan dimplementasikan untuk melatih dan mengetes dataset sesuai dengan pembagian sebelumnya. Adapun cara kerja Random Forest adalah :

- Penentuan jumlah *Base Learner (Decision Tree)*
- Mencari hasil dari jetiap *Base Learner (Decision Tree)*
- Pencarian hasil prediksi dengan melakukan akumulasi hasil dari setiap *Base Learner (Decision Tree)* dengan *Majority Voting*

2.4. Pengujian Model dan Evaluasi

Setelah beberapa tahapan diatas dilakukan, model sudah bisa ditesting menggunakan data baru untuk mendeteksi penyakit jantung koroner pada seseorang dengan menginputkan data-data berdasarkan variable-variabel sebelumnya. Setelah algoritma Random Forest diimplementasikan, selanjutnya

p-ISSN: 2301-5373
e-ISSN: 2654-5101

peneliti menampilkan classification report dari algoritma Random Forest itu sendiri, yang berisi akurasi model, training score model, matrix confusion dan cross validation score dari model.

Gambar 1 menjelaskan alur diagram dari metode penelitian yang dilakukan.

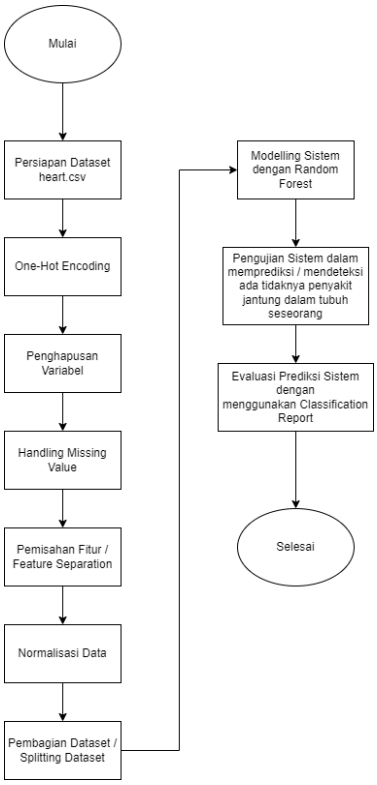

Figure 1. Metode Penelitian

3. Hasil dan Pembahasan

Pada penelitian ini, peneliti akan membuat program komputer untuk mengimplementasikan teknik, metode dan algoritma yang akan digunakan pada sistem deteksi. Peneliti menggunakan bahasa Python dalam pengembangan sistem ini peneliti menggunakan Jupyter untuk membuat kode program. Pengembangan sistem ini menggunakan algoritma Random Forest dengan memanfaatkan library pada python yaitu sklearn.

3.1. Dataset

Dataset yang digunakan adalah dataset dengan nama heart.csv yang diakses diakses dari situs Kaggle.com yaitu https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset yang didalamnya berisi 303 baris dengan 14 kolom yang merupakan variabel yang terdiri atas 9 variabel kategorik dan 5 variabel kontinu. 9 variabel kategorik tersebut adalah sex, cp, fbs, restecg, exang, slope, ca, thal, target. Sedangkan 5 variabel kontinu tersebut adalah age, trestbsp, chol, thalach dan oldpeak. Ke 14 variabel inilah yang merupakan data-data Kesehatan yang berhubungan dengan jantung. Dibawah ini adalah gambar dataset heart.csv :

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

Figure 2. Dataset heart.csv

3.2. Preprocessing Data

Seperti yang telah dibahas sebelumnya, bahwa data perlu dikelola dengan dilakukan preprocessing sebelum digunakan dalam sistem.

a. One-Hot Encoding

Diterapkan pada variable kategorik yang lebih dari 2 kategori (bukan binary) yaitu variable 'cp', 'thal', 'slope'. One-hot encoding diperlukan sehingga data kategori yang bukan binary tersebut bisa diubah ke dalam bentuk binary sehingga mudah untuk diproses pada sistem. Data frame yang terbentuk setelah dilakukan one-hot encoding dapat dilihat pada table dibawah ini :

	age	sex	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	ca	...	cp_1	cp_2	cp_3	thal_0	thal_1	thal_2	thal_3	slope_0	slope_1	slope_2
0	63	1	145	233	1	0	150	0	2.3	0	...	0	0	1	0	1	0	0	1	0	0
1	37	1	130	250	0	1	187	0	3.5	0	...	0	1	0	0	0	1	0	1	0	0
2	41	0	130	204	0	0	172	0	1.4	0	...	1	0	0	0	0	1	0	0	0	1
3	56	1	120	236	0	1	178	0	0.8	0	...	1	0	0	0	0	1	0	0	0	1
4	57	0	120	354	0	1	163	1	0.6	0	...	0	0	0	0	0	1	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	140	241	0	1	123	1	0.2	0	...	0	0	0	0	0	0	1	0	1	0
299	45	1	110	264	0	1	132	0	1.2	0	...	0	0	1	0	0	0	1	0	1	0
300	68	1	144	193	1	1	141	0	3.4	2	...	0	0	0	0	0	0	1	0	1	0
301	57	1	130	131	0	1	115	1	1.2	1	...	0	0	0	0	0	0	1	0	1	0
302	57	0	130	236	0	0	174	0	0.0	1	...	1	0	0	0	0	1	0	0	1	0

303 rows × 22 columns

Figure 3. Dataset heart.csv setelah dilakukan One-Hot Encoding

b. Penghapusan Variabel

Masih berkaitan dengan One-Hot Encoding, dimana variable 'cp', 'thal', 'slope' sudah diubah dengan One-Hot Encoding, yang mana menjadi ‘cp_1’, ‘cp_2’, ‘cp_3’, ‘thal_1’, ‘thal_2’, ‘thal_3’, ‘slope_1’, ‘slope_2’, ‘slope_3’, maka ketiga variable asli ini (‘cp’, ‘thal’, ‘slope’) sudah tidak diperlukan lagi.

c. Handling Missing Values

Missing value adalah nilai yang tidak terdefinisi di dataset. Missing value perlu ditangani karena dapat menimbulkan perubahan hasil analisis. Pada akhirnya, data-data yang memuatnya bisa memberikan kesimpulan yang berbeda dibandingkan dengan data yang telah dibersihkan atau dibenahi. Pada penelitian ini, peneliti menghapus missing values sehingga tidak terjadi kesalahan interpretasi.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

```
age         0
sex         0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
ca          0
target      0
cp_0        0
cp_1        0
cp_2        0
cp_3        0
thal_0      0
thal_1      0
thal_2      0
thal_3      0
slope_0     0
slope_1     0
slope_2     0
dtype: int64
```

Figure 4. Variabel yang memiliki missing values

d. Pemisahan Fitur / Feature Separation

Pemisahan fitur dilakukan untuk memisahkan fitur / variable x dari variable y sehingga bisa diproses masing-masing. Hal ini dilakukan agar sistem mengenal variable mana yang akan diguanakn sebagai inputan untuk mendapatkan hasil prediksi dan variable mana yang dikatakan sebagai hasil prediksi. Pada penelitian ini, variable x adalah variable yang digunakan untuk inputan dalam mendeteksi penyakit jantung koroner, dan variable y adalah hasil prediksinya.

e. Normalisasi Data

Pada bagian ini, akan dilakukan normalisasi data untuk menormalkan rentang variabel independen atau fitur data untuk alur kerja data yang lebih baik. Normalisasi data akan menggunakan normalisasi min-max. Normalisasi min-max sering dikenal sebagai penskalaan fitur di mana nilai rentang numerik fitur data, dikurangi menjadi skala antara 0 dan 1.

f. Pembagian Dataset / Splitting Dataset

Pada penelitian ini, peneliti membagi dataset menjadi 2 bagian yaitu 80% data training dan 20% data testing.

g. Modelling Sistem dengan Random Forest

Random Forest bekerja dengan 3 tahap yaitu penentuan jumlah *base learner (Decision Tree),* mencari hasil dari setiap *base learner (Decision Tree) dan* pencarian hasil prediksi dengan melakukan akumulasi hasil dari setiap *base learner (Decision Tree)* dengan *Majority Voting.* Dalam penelitian ini, peneliti memanfaatkan library python yang bernama sklearn untuk melakukan 3 tahapan ini, dimana implementasi algoritma Random Forest secara keseluruhan sudah bisa dilakukan / diimplementasikan oleh library ini.

h. Pengujian dan Evaluasi Sistem

Setelah semua persiapan dan implementasi algoritma Random Forest selesai, maka dilanjutnya dengan melakukan pengujian sistem untuk melihat apakah sistem sudah bekerja dengan baik dengan cara mengukur akurasi model, training score model, matrix confusion dan cross validation score Pada penelitian ini, model memiliki akurasi sebesar 91%, training score dan cross validation yang memiliki kurva naik yang menunjukan bahwa sistem dapat belajar dengan baik dari dataset yang digunakan dan juga perhitungan kebenaran prediksi yang tinggi dibandingkan perhitungan kesalahan prediksi yang didapatkan dari confusion matrix yakni bernilai 24 : 4. Hasil evaluasi model ini dapat dilihat melalui classification report dan table evaluasi yang dapat dilihat pada gambar dibawah ini :

```
.:. Random Forest Accuracy: 91.80% .:.

.: Classification Report
************************
              precision    recall  f1-score   support

           0       0.86      0.96      0.91        25
           1       0.97      0.89      0.93        36

    accuracy                           0.92        61
   macro avg       0.91      0.92      0.92        61
weighted avg       0.92      0.92      0.92        61
```

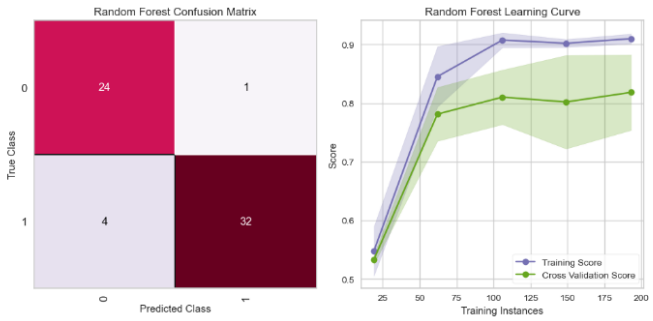

Figure 5. Gambar Classification Report

Table 1. Tabel hasil akurasi algoritma Random Forest

Model	Akurasi
Random Forest	91.803279%

Berdasarkan classification report diatas dan table akurasi diatas, dapat dikatakan bahwa sistem sudah berjalan dengan baik, maka dari itu, kita sudah bisa melakukan pengujian sistem. Pengujian dilakukan dengan menginputkan data-data 13 variabel y dari seseorang, yaitu “age”, “sex”, “trestbps”, “chol”, “fbs”, “restecg”, “thalach”, “exang”, “oldpeak”, “ca”, “cp”, “thal” dan “slope”. Jika hasil prediksi adalah 1 maka penyakit jantung koroner dikatakan terdeteksi, sedangkan jika hasil prediksi adalah 0 maka penyakit jantung koroner tidak terdeteksi, yang dapat dilihat pada gambar dibawah ini :

```
# --- Input Data Pasien ---
data = [[0.254, 1, 0.487, 0.362,   ## age_scaled, sex, trestbps_scaled, chol
         1, 0.5, 0.641, 1,         ## fbs, restecg_scaled, thalach_scaled, exang
         0.672, 0.863, 0, 0,       ## oldpeak_scaled, ca_scaled, cp_0, cp_1
         0, 1, 0, 0,               ## cp_2, cp_3, thal_0, thal_1
         0, 1, 0, 1, 0]]          ## thal_2, thal_3, slope_0, slope_1, slope_2

# --- Prediksi dari algoritma Random Forest ---
result = RFclassifier.predict(data)

# --- Cetak status penyakit jantung koroner pasien ---
if result[0] == 1:
  print('\033[1m' + '.:. Penyakit Jantung Koroner Terdekteksi!.:.' + '\033[0m')
elif result[0] == 0:
  print('\033[1m' + '.:. Penyakit Jantung Koroner Tidak Terdekteksi!.:.' + '\033[0m')
```

```
.:. Penyakit Jantung Koroner Tidak Terdekteksi!.:.
```

Figure 9. Gambar Pengujian Sistem

4. Conclusion

Dari penelitian ini, dapat disimpulkan bahwa algortima Random Forest sudah dapat dikatakan sebagai algoritma yang cukup akurat dalam memprediksi penyakit jantung koroner, karena memiliki classification report yang cukup baik yaitu akurasi sebesar 91%, training score dan cross validation yang memiliki kurva naik dan juga perhitungan akurasi yang cukup baik dari confusion matrix. Dan dapat disimpulakan pula bahwa algoritma Random Forest sudah cocok dalam melakukan pendeteksian, dalam hal ini mendeteksi penyakit jantung koroner.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

References

[1] K. Sutomo, “Gangguan Metabolisme Lemak dan Penyakit Jantung Koroner,” *Pidato Pengukuhan Jabatan Guru Besar Tetap dalam Ilmu Penyakit Dalam pada Universitas Sumatra Utara. Medan*, 2019.

[2] B. Sadikin, “Keputusan Menteri Kesehatan Republik Indonesia Tentang Pedoman Penyakit Jantung dan Pembuluh Darah,” *World Health Organization*, no. 854/MENKES/SK/IX/2009, 2009.

[3] B. Sadikin, “Penyakit Tidak Menular (PTM) Penyebab Kematian Terbanyak Di Indonesia,” *Kementerian Kesehatan Republic Indonesia,* 2011.

[4] Larose, “Discovering Knowledge in Data,” New Jersey : John Willey, 2013.

[5] W. Frank, H, ”Data Mining : Practical Machine Learning and Tools,” *Morgan Kaufmann Publisher*, 2011.

This page is intentionally left blank