ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

Predicting Heart Disease Using XgBoost Algorithm and RandomizedSearch Optimizer

Prediksi Penyakit Jantung dengan Menggunakan Algoritma XgBoost dan RandomizedSearch Optimizer

Reo Sahobby[1], Dessyanto Boedi P[2], Mangaras Yanu F[3]

[1,2,3] Informatika, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia

[1*]123170067@student.upnyk.ac.id, [2]dess@upnyk.ac.id, [3]mangaras.yanu@upnyk.ac.id
**: Penulis korenspondensi* (*corresponding author*)

Informasi Artikel
Received: December 2021
Revised: -
Accepted: -
Published: -

Abstract

Purpose: The purpose of this research is to predict whether there is a possibility of heart disease using the XgBoost algorithm based on tabular data or table data that has 13 parameters such as age, gender, blood pressure, and others. Another goal is to reduce overfitting, because the XgBoost algorithm is known to have the advantage of being able to suppress overfitting but still maintain high accuracy.
Design/methodology/approach: The design and method used in this research is to carry out data preprocessing processes such as data cleaning, data splitting, and data transformation according to data types. After doing the preprocessing process, a training process will be carried out for making prediction models. In making the prediction model, to maximize the XgBoost algorithm, the Randomized Search Optimizer is used to get the best hyper parameters so that it can produce a model with the best possible accuracy.
Findings/result: Machine learning models created with the XgBoost algorithm get 91% accuracy on training data and 83% on testing data. Tests carried out using other datasets get 78% accuracy, and the general model gets 90% accuracy.
Originality/value/state of the art: This research was conducted using the XgBoost algorithm combined with the RandomizedSearch Optimizer as a hyper parameter tuning for machine learning model making.

Keywords:machine learning; classification; heart diseases
Kata kunci: pembelajaran mesin; klasifikasi; penyakit jantung

Abstrak

Tujuan: Tujuan dari penelitian ini adalah untuk memperdiksi ada tidaknya kemungkinan penyakit jantung menggunakan algoritma XgBoost berdasarkan data *tabular* atau data tabel yang memiliki 13 parameter seperti umur, jenis kelamin, tekanan darah, dan lain-lain. Tujuan lainnya adalah untuk mengurangi *overfitting*, karena algoritma XgBoost dikenal memiliki kelebihan dapat menekan *overfitting* namun tetap mempertahankan akurasi yang tinggi.

Perancangan/metode/pendekatan: Perancangan dan metode yang digunakan dalam penelitian ini adalah melakukan proses *preprocessing* data seperti data *cleaning*, data *splitting*, dan data *transformation* sesuai dengan tipe data. Setelah melakukan proses *preprocessing*, akan dilakukan proses *training* untuk pembuatan model prediksi. Dalam pembuatan model prediksi, untuk memaksimalkan algoritma XgBoost maka digunakan Randomized Search Optimizer untuk mendapatkan *hyper parameter* terbaik sehingga dapat menghasilkan model dengan akurasi sebaik mungkin.

Hasil: Model *machine learning* yang dibuat dengan algoritma XgBoost mendapatkan akuasi 91% pada *data training* dan 83% pada *data testing*. Pengujian yang dilakukan menggunakan *dataset* lain mendapatkan akurasi 78%, dan model yang dibuat secara umum mendapatkan akuasi sebesar 90%.

Keaslian/ *state of the art*: Penelitian ini dilakukan menggunakan algoritma XgBoost yang dikombinasikan dengan RandomizedSearch Optimizer sebagai tuning *hyper parameter* untuk pembuatan model *machine learning*.

1. Pendahuluan

Jantung merupakan organ dalam manusia yang memiliki fungsi sangat penting yaitu untuk mengedarkan darah yang berisi oksigen dan nutrisi ke seluruh tubuh dan untuk mengangkut sisa hasil metabolisme tubuh, sehingga tubuh dapat bekerja dengan optimal. Sehingga akan sangat fatal apabila di dalam organ jantung terdapat gangguan seperti penyumbatan pembuluh darah, dan lain-lain. Penyakit jantung adalah penyakit yang menyerang organ jantung, contohnya adalah penyumbatan pembuluh darah pada jantung. Penyakit ini menyerang pembuluh darah arteri karena terjadi proses *arteosklerosis* pada dinding *arteri* yang menyebabkan penyempitan [1]**.**

Penyakit jantung dapat disebabkan oleh beberapa faktor seperti peningkatan kadar kolesterol karena dapat menyebabkan penumpukan lemak pada dinding arteri dan dapat menyebabkan *arteriosklerotik*. Selain itu dapat juga disebabkan oleh peningkatan tekanan darah atau *hipertensi*, karena saat tekanan darah meningkat maka dapat membebani kerja jantung dan juga menyebabkan *arteriosklerotik* karena saat tekanan darah meningkat akan menyebabkan gaya renggang yang dapat merobek lapisan *endotel arteri* dan *arteriol*. Kemudian penyakit jantung juga dapat disebabkan oleh kebiasaan merokok, orang dengan kebiasaan merokok mempunyai risiko 2,3 kali lebih besar untuk terkena penyakit jantung pada usia kurang dari 45 tahun.

Gejala klinis penyakit jantung antara lain adalah sering merasakan sesak napas yang ditandai dengan napas yang berat dan pendek sewaktu melakukan aktivitas berat, semakin lama rasa sesak napas akan semakin bertambah. Gejala lainnya adalah *klaudiokasi intermiten*, yaitu rasa nyeri pada daerah ekstrimitas bawah dan terjadi selama atau setelah melakukan olah raga. Gejala lainnya adalah mengalami perubahan warna kulit, dan kadar kolesterol yang meningkat biasanya di atas 180mg/dl untuk usia kurang dari 30 tahun, dan di atas 200mg/dl untuk yang berusia lebih dari 30 tahun.

Melihat dari tingkat bahaya penyakit jantung, maka penelitian yang membahas tentang penyakit jantung sudah banyak dilalukan menggunakan beberapa algoritma, namun dari beberapa penelitian yang ada dirasa masih memiliki kekurangan yaitu *overfitting* yang ada pada model yang dibuat. *Overfitting* adalah kondisi dimana model *machine learning* yang dibuat memiliki hasil akurasi pada *data training* yang sangat bagus, namun pada saat dilakukan pengujian menggunakan *data testing* didapatkan hasil akurasi model yang buruk [2].

Beberapa penelitian yang sudah dilakukan diantaranya adalah penelitian dengan judul Peningkatan Akurasi Klasifikasi Algoritma C4.5 Menggunakan Teknik Bagging Pada Diagnosis Penyakit Jantung, penelitian tersebut dilakukan untuk memprediksi penyakit jantung sekaligus untuk meningkatkan akurasi algoritma C4.5 apabila dikombinasikan dengan teknik *bagging*. Data yang digunakan dalam penelitian tersebut adalah data *cleveland dataset* yang terdiri dari 303 data dan memiliki 13 parameter. Untuk melakukan validasi performa algoritma dilakukan menggunakan *k-fold cross validation* dengan k=10 dan, untuk mengukur akurasi algoritma dalam penelitian tersebut dilakukan menggunakan *confusion matrix*. Hasil dari penelitian tersebut adalah model yang dibuat menggunakan algoritma C4.5 mendapatkan akurasi sebesar 72,98% sedangkan model yang dikombinasikan dengan teknik *bagging* mendapatkan akurasi sebesar 81,84%. Dapat dikatakan bahwa teknik *bagging* yang dikombinasikan dengan algoritma C4.5 dapat meningkatkan akurasi sebesar 8,86% [3]. Kemudian penelitian selanjutnya dengan judul Prediksi Penyakit Jantung Dengan Algoritma Klasifikasi, penelitian ini dilakukan untuk mengetahui perbandingkan akurasi model yang dibuat dengan beberapa algoritma seperti *Naive Bayes*, SVM, C4.5, *Logistic Regression*, dan *Back Propagation*. *Dataset* yang dilakukan dalam penelitian tersebut adalah *statelog heart diseases* yang berjumlah 270 data dan 13 parameter. Proses *training* yang dilakukan sekaligus validasi yang dilakukan dengan *cross validation*. Hasil dari penelitian tersebut adalah model yang dibuat dengan algoritma *Naive Bayes* mendapat akurasi paling tinggi sebesar 84,07%. Untuk pengukuran presisi paling tinggi adalah algoritma *Naive Bayes* dengan 86,16% sedangkan *recall* paling tinggi adalah algoritma SVM dengan *recall* mencapai 94,67% [4]. Selanjutnya penelitian dengan judul Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung, penelitian tersebut dilakukan untuk membandingkan

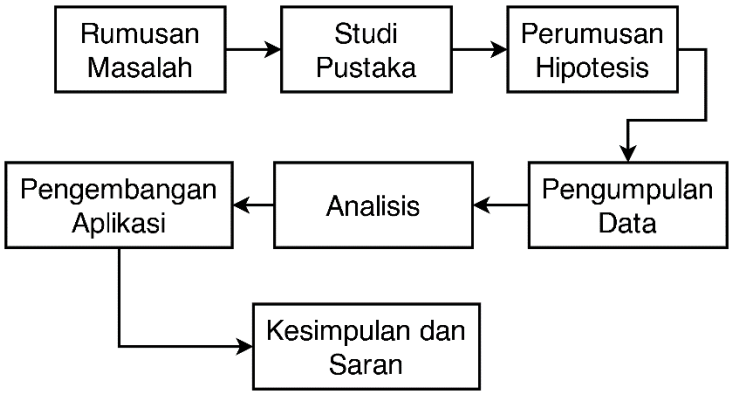
ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

beberapa algoritma dalam melakukan klasifikasi penyakit jantung. Algoritma yang dilakukan dalam penelitian tersebut diantaranya adalah *Naive Bayes*, KNN, *Decison Tree*, *Random Forest*, dan SVM. Data yang digunakan dalam penelitian tersebut adalah *cleveland heart disesase* yang berjumlah 303 data dengan 13 parameter, dalam penelitian tersebut dilakukan *dataset splitting* dengan perbaingkan 80% sebagai *data training* dan 20% sebagai *data testing*, juga dilakukan validasi menggunakan *cross validation* dengan nilai k=5. Dari proses *training* dan pengujian yang sudah dilakukan, didapatkan hasil akurasi paling tinggi didapatkan pada algoritma *Random Forest* dengan 85,67%, algoritma *Naive Bayes* dan *Decision Tree* mendapat akurasi sebesar 69,67%. Dari penelitian tersebut dapat dsimpulkan bahwa algoritma *Random Forest* memiliki performa yang paling baik dalam menangani permasalahan klasifikasi penyakit jantung [5].

Dari penjelasan dan penelitian sebelumnya yang sudah ada, penelitian ini dilakukan untuk memprediksi penyakit jantung menggunakan algoritma XgBoost sekaligus mengurangi *overfitting* karena algoritma XgBoost memiliki kelebihan dapat mencegah *overfitting* dengan *regularization* sebagai teknik untuk mencegah terjadinya *overfitting*. Hasil dari penelitian ini adalah model *machine learning* dan *confusion matrix* untuk mengetahui akurasi algoritma XgBoost dalam membuat model prediksi penyakit jantung.

2. Metode/Perancangan

Metode penelitian yang dilakukan ini adalah penelitian kuantitatif, untuk tahapan penelitian ini dapat dilihat pada **Gambar 1.**



Gambar 1. Tahapan Penelitian

Seperti yang dapat dilihat pada **Gambar 1.** penelitian ini terdari dari beberapa tahapan seperti perumusan masalah, studi pustaka, hingga pengembangan aplikasi. Namun dari tahapan yang ada, akan dilakukan pembahasan lebih pada tahapan analisis dan pengembangan aplikasi.

2.1. Pengumpulan Data

Data yang digunakan dalam pelenitian ini adalah data sekunder, data sekunder adalah data yang bisa didapatkan secara tidak langsung misalnya melalui orang lain, sumber dokumen, *website*, dan sumber lainnya [6]. Data yang digunakan adalah *cleveland dataset* yang berjumlah 303 data dengan pembagian target 138 data tergolong dalam klasifikasi tidak memiliki penyakit jantung, dan 165 data tergolong dalam klasifikasi memiliki penyakit jantung. Data *cleveland dataset* memiliki 13 parameter yang dapat dilihat pada **Tabel 1.**

Tabel 1. Parameter Dataset

No.	Parameter	Deskripsi	Keterangan
1.	Age	Umur pasien	numerik
2.	Sex	Jenis kelamin pasien	0: wanita, 1: pria
3.	Cp	Chest pain type	1: typical angina, 2: atypical angina 3: non-angina pain, 4: asympotomatic
4.	Trestbps	Tekanan darah pasien	Numerik
5.	Chol	Kadar serum kolesterol	Numerik
6.	Fbs	Kadar gula darah apakah > 120mg/dl	0: false, 1: True
7.	Restecg	Hasil ECG selama istirahat	0: normal, 1: abnormal (memiliki kelainan pada gelombang ST-T) 2: hipertrofil ventrikel
8.	Thalach	Detak jantung maksimal yang dicapai	numerik
9.	Exang	Ukuran boolean yang menunjukan apakah latihan angina terjadi	0: no 1: yes
10.	Oldpeak	Segment ST yang diperoleh dari hasil ECG	Numerik
11.	Slope	Jenis kemiringan segment ST untuk latihan maksimal (puncak)	1: upsloping 2: flat, 3: downsloping
12.	Ca	Jumlah vessel yang diwarnai oleh fluroskopi	0, 1, 2, 3
14.	Thal	Parameter thalasemia	1: normal, 2: cacat tetap, 3: reversible
14.	Target	Target kelas klasifikasi	0: tidak terkena penyakit jantung 1: terkena penyakit jantung

Seperti yang dapat dilihat pada **Tabel 1.** jumlah parameter adalah 13 dan 1 kolom target yang terdiri dari 0 dan 1. Untuk penjelas lebih mengenai parameter adalah sebagai berikut.

- Cp (*Chest Pain Type*)
  Cp adalah nyeri dada yang disebabkan karena otot jantung tidak mendapatkan cukup darah yang kaya dengan oksigen. Nyeri dada dapat menjadi kekhawatiran pada masalah jantung, sehingga sangan umum ketika gejala ini muncul orang akan mengira terjadi serangan jantung.
- Trestbps
  Tekanan darah adalah ukuran kekuatan yang digunakan jantung untuk memompa darah ke seluruh tubuh. Pengaruh dalam penyakit jantung adalah apabila tekanan darah mengalami pengingkatan. Pengingkatan tekanan darah merupakan beban yang berat sehingga akan menyebabkan *hipertropi ventrikel* atau pembesaran *ventrikel*.
- Chol
  Serum kolesterol adalah senyawa lipid yang mempunyai inti *siklopenta perhidrofenanta* [7]. Permasalahan dalam penyakit jantung adalah apabila terjadi penginkatan kadar serum kolesterol. Ukuran kolesterol normal sebaiknya kurang dari 200mg/dl.
- Fbs
  Fbs adalah ukuran kadar gula yang dilakukan pasien setelah berpuasa selama 8 jam. Ukuran fbs apabila melebihi 120mg/dl maka dapat dikatakan bahawa pasien tersebut memiliki potensi untuk *diabetes*. Pengaruh *diabetes militus* dalam kesehatan jantung

adalah pasien yang memiliki *diabetes militus* mempunyai peluang 10,25 kali lebih besar untuk terkena penyakit jantung daripada pasien yang tidak memiliki *diabetes militus*.

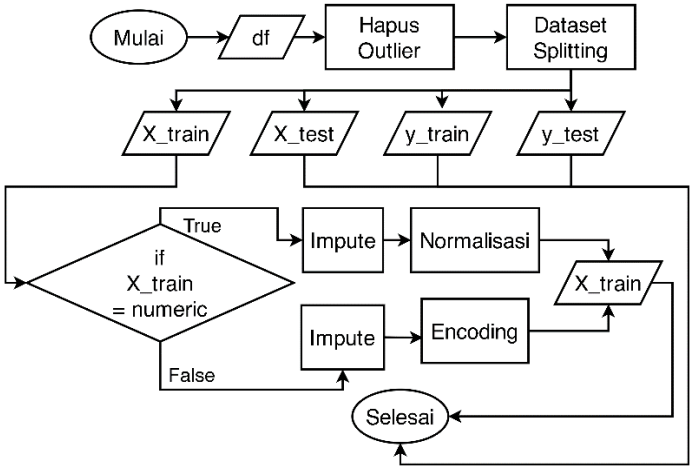
- Restecg
  Restecg adalah hasil pengukuran ECG pada jantung. ECG dilakukan dengan cara menempatkan sepuluh elektroda pada titik-titik tertentu, enam elektroda dipasangkan di area dada, dan selebihnya dipasangkan pada area ekstrimitas. Pemeriksaan ECG merupakan hal yang wajib dilakukan kepada pasien yang memiliki tanda-tanda atau gejala penyakit jantung [8].
- Thalach
  Thalach adalah ukuran detak jantung maksimal yang dapat dicapai. Detak jantung manusia normal berkisar antara 60 – 100 kali per menit [9].
- Exang
  Exang adalah informasi mengenai latihan angina (nyeri dada) yang dilakukan, apabila selama aktivitas latihan angina dilakukan pasien merasakan rasa sakit maka *exang* akan bernilai 1, dan apabila tidak merasakan rasa sakit maka *exang* akan bernilai 0 [10].
- Oldpeak
  Oldpeak adalah ukuran gelombang segment ST yang terjadi diakibatkan oleh latihan *relative* terhadap kondisi jantung saat beristirahat atau tidak bekerja keras. Ukuran oldpeak biasanya didefinisikan dengan nilai 0 sampai 3mm [11].
- Slope
  Slope adalah jenis kemiringan segment ST yang terdapat pada gambar grafik hasil pemerikasaan ECG, segment ST yang dihasilkan dari pemeriksaan nantinya dapat dihubungkan dengan ketidaknormalan pada dinding jantung [12].
- Ca
  Ca adalah jumlah dari *vessel* yang diwarnai oleh *fluroskopi*. Teknik *fluroskopi* adalah teknik yang memanfaatkan salah satu sinar X yaitu jika sinar tersebut terkena bahan maka akan berpenda menjadi warna tertentu [13]. *Fluroskopi* juga dapat digunakan untuk menunjang prosedur medis tertentu contohnya pada prosedur pemasangan *ring* jantung [14].
- Thal
  Thal adalah parameter *thalasemia* yang dimiliki pasien. *Thalasemia* adalah sebuah penyakit keturunan yang diakibatkan oleh gagalnya pembentukan satu dari empat asam amino yang membentuk *hemoglobin*, sehingga *hemoglobin* tidak terbentuk secara sempurna [15].

2.2. Analisis

2.2.1. Preprocessing

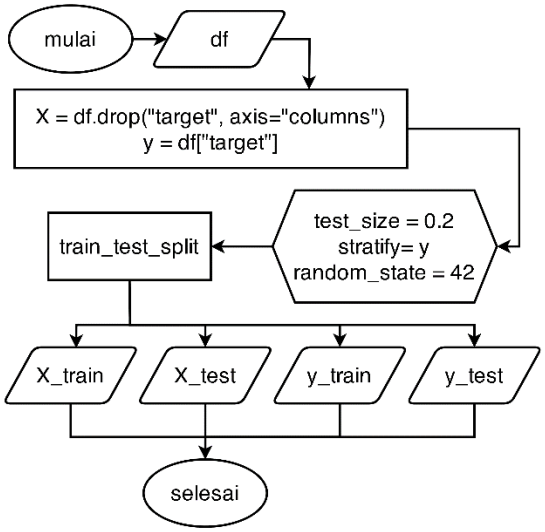
Tahapan proses *preprocessing* yang dilakukan pada penelitian ini dapat dilihat pada *Flowchart* yang ditampilkan pada **Gambar 2.**

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx



Gambar 2. Flowchart Preprocessing

Seperti yang dapa dilihat pada **Gambar 2.** proses *preprocessing* yang dilakukan terdiri dari beberapa proses seperti hapus *outlier*, kemudian melakukan *dataset splitting* untuk membagi data menjadi *data training* dan *data testing*. Flowchart dataset splitting dapat dilihat pada **Gambar 3.**



Gambar 3. Flowchart Dataset Splitting

Dapat dilihat pada **Gambar 3.** bahwa perbandingan data hasil dari proses *dataset splitting* adalah 80% dari *dataset* akan menjadi *data training* yang akan digunakan untuk proses pembuatan model, dan 20% dari *dataset* yang ada akan menjadi *data testing* yang nantinya akan digunakan untuk pengujian dan pengukuran akurasi model yang sudah dibuat. Proses *dataset splitting* akan menghasilkan empat *dataframe* baru, dan pada *dataframe* X_train akan dilakukan proses *preprocessing* berikutnya sesuai dengan tipe data. Pada data dengan tipe data numerik akan dilakukan proses normalisasi, proses normalisasi dilakukan supaya proses *training* menjadi lebih cepat karena dapat memudahkan model untuk memahami data [16]. Sedangkan pada data dengan tipe data kategorik akan dilakukan proses *encoding* yaitu proses merubah data

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

kategori dalam satu kolom menjadi data *binnary* dan jumlah kolom akan bertambah sesuai dengan kategori yang ada.

2.2.2. Training

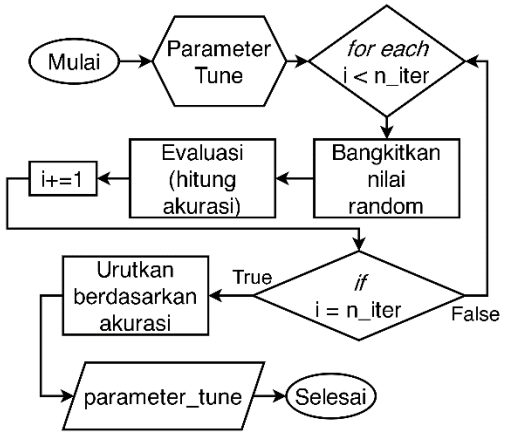
Proses pembuatan model atau *training* dalam penelitian ini dilakukan menggunakan algoritma XgBoost, algoritma XgBoost membuat beberapa *tree* sekaligus dan mempunyai *lambda regularization* sebagai teknik untuk mencegah terjadinya *overfitting*.Algoritma XgBoost memiliki rumus utama untuk menentukan nilai *similarity*, nilai *gain*, dan *output value* yang dapat dilihat pada **persamaan 1**, **2**, dan **3.**

$$similarity = \frac{\sum(residual)^2}{\sum[prev\ probabillity_i \times (1 - prev\ probabillity_i)] + \lambda} \quad (1)$$

$$gain = left\ similarity + right\ similarity - root\ similarity \quad (2)$$

$$O_{value} = \frac{\sum(residual_i)}{\sum[prev\ probabillity_i \times (prev\ probability_i)] + \lambda} \quad (3)$$

Selanjutnya, proses *training* dalam penelitian ini juga dibantu menggunakan *Randomiezed Search* yang berfungsi untuk menentukan *hyper parameter* terbaik sehingga dapat membuat model dengan akurasi yang paling tinggi. Flowchart dari Randomized Search dapat dilihat pada **Gambar 4.**



Gambar 4. Flowchart Randomized Search Optimizer

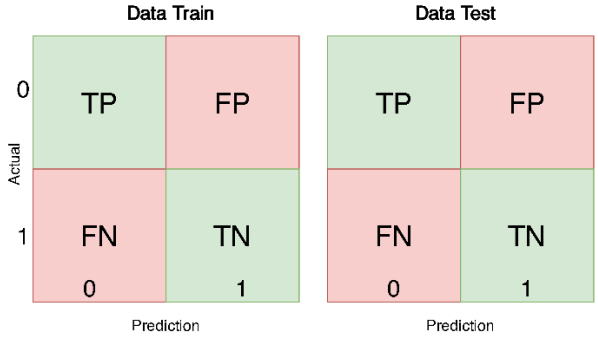
Dapat dilihat pada **Gambar 4.** proses penentuan *hyper parameter* terbaik model dilakukan dengan *Randomized Search*. Proses diawali dengan menentukan range *hyper parameter* apa saja yang akan dituning dan berapa jumlah iterasi yang diinginkan, kemudian *Randomized Search* akan membuat kombinasi nilai *random* masing-masing *hyper parameter* sebanyak iterasi yang telah ditentukan. Untuk masing-masing iterasi *Randomized Search* akan membuat model dan melakukan pengujian dengan *data testing*, kombinasi *hyper parameter* yang dapat membuat model dengan akurasi paling tinggi akan dijadikan parameter model.

2.2.3. Evaluasi

Evalusai yang dilakukan bertujuan untuk melihat performa model *machine learning* yang dibuat. Proses evaluasi dilakukan dengan cara menguji model yang sudah dibuat menggunakan *data train* dan *data testing* yang ada, hasil proses *dataset splitting* dan *preprocessing* yang sudah dijelaskan di atas. Proses evaluasi juga dilakukan dengan *dataset* lain yaitu *statelog dataset* yang

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

berjumlah 270 data namun tetap memiliki parameter data yang sama yaitu 13 parameter. Hasil dari proses evaluasi ditampilkan dalam bentuk *confussion matrix*, sehingga dapat menghitung akurasi model dari *confussion matrix* tersebut. Gambar dari rancangan *confusion matrix* dapat dilihat pada **Gambar 5.**



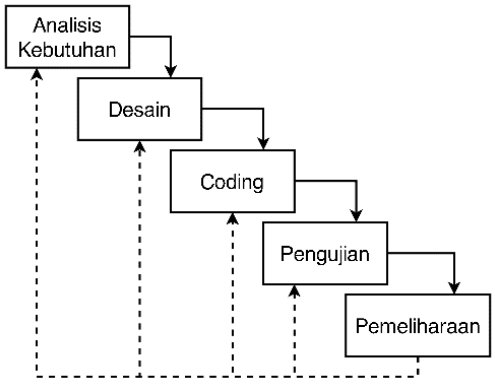
Gambar 5. Rancangan Confusion Matrix

Seperti yang dapat dilihat pada **Gambar 5.** *confusion matrix* yang digunakan dalam evalusai penelitian ini akan dibagi menjadi dua masing-masing sesuai dengan *data training* dan *data testing*, sehingga dapat digunakan untuk mengukur perbandingan dan selisih antara akurasi pada *data training* dan akurasi pada *data testing*. Dalam *confusion matrix* terdapat empat bagian diantaranya TP yang merupakan fakta benar yang dapat diprediksi benar, FP merupakan fakta salah yang diprediksi benar, FN yang merupakan fakta benar yang diprediksi salah, dan TN yang merupakan fakta salah yang diprediksi salah. Untuk mengukur akurasi berdasarkan confusion matrix dapat menggunakan rumus yang ada pada **Persamaan 4.**

$$akurasi = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

2.3. Pengembangan Aplikasi

Pengembangan aplikasi yang dilakukan dalam penelitian ini dilakukan dengan model pengembangan *waterfall*. Untuk tahapan *waterfall* dapat dilihat pada **Gambar 6.**



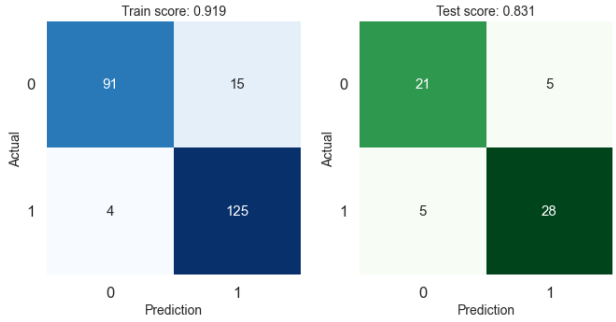
Gambar 6. Tahapan Waterfall

Seperti yang dapat dilihat pada **Gambar 6.** tahapan *waterfall* diawali dari analisis kebutuhan baik kebutuhan fungsional maupun non fungsional, kemudian desain, coding, pengujian aplikasi, hingga pemeliharaan aplikasi setelah selesai dibuat.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

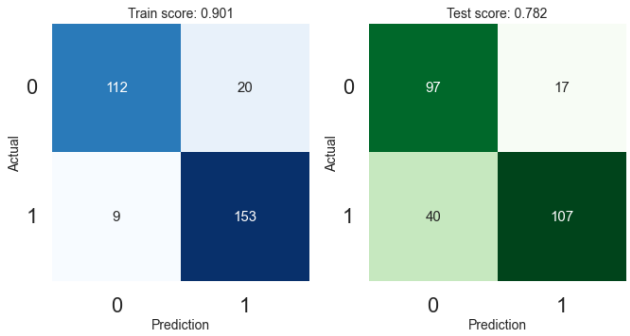
3. Hasil dan Pembahasan

Hasil dari penelitian ini adalah model *machine learning* yang sudah dibuat menggunakan algoritma XgBoost dan *Randomized Search Optimizer*. Untuk mengetahui hasil evaluasi dan akurasi model dapat dilakukan menggunakan *confusion matrix* yang dapat dilihat pada **Gambar 7.**



Gambar 7. Confusion Matrix Model

Pada **Gambar 7.** adalah *confusion matrix* hasil pengujian yang dilakukan menggunakan *data training* dan *data testing* hasil dari *dataset splitting* dengan perbandingan 80% *data training* dan 20% *data testing*. Dari **Gambar 7.** yang dapat dilihat, model yang dibuat mendapatkan akurasi sebesar 91% pada pengujian menggunakan *data training*, dan akurasi sebesar 83% pada pengujian yang dilakukan menggunakan *data testing*. Proses evaluasi dan pengujian juga dilakukan menggunakan *dataset* lain yaitu *dataset statelog heart diseases* yang berjumlah 270 data, *confusion matrix* untuk pengujian dengan *dataset statelog* dapat dilihat pada **Gambar 8.**



Gambar 8. Confusion Matrix Pada Dataset Lain

Seperti yang dapat dilihat pada **Gambar 8.** *confusin matrix* yang ditampilkan dibagi menjadi dua yaitu sebelah kiri adalah *confusion matrix* hasil pengujian pada seluruh *dataset* yang digunakan yaitu *cleveland heart diseases*, dan sebelah kanan adalah *confusion matrix* hasil pengujian yang dilakukan menggunakan *dataset* lain yaitu menggunakan *statelog heart diseases*. Dapat dilihat pada *confusion matrix* sebelah kiri bahwa model yang dibuat mendapatkan akurasi 90% apabila dilakukan pengujian menggunakan seluruh *dataset cleveland heart diseses*. Sedangkan pada *confusin matrix* sebelah kanan dapat dilihat bahwa model yang dibuat mendapat akurasi sebesar 78% apabila dilakukan pengujian menggunakan *dataset* lain yaitu *statelog heart diseases*.

4. Kesimpulan dan Saran

Berdasarkan penelitian yang sudah dilakukan dah hasil penelitian yang sudah didapatkan, dapat disimpulkan bahwa.

- algoritma XgBoost dapat digunakan untuk menyelesaikan permasalahan prediksi penyakit jantung menggunakan data *tabular* atau data tabel yang terdiri dari 13 parameter. Model *machine learning* yang dibuat mendapatkan akurasi sebesar 90% dengan rincian akurasi pada *data training* sebesar 91% dan akurasi pada *data testing* sebesar 83%.
- Algoritma XgBoost yang digunakan dalam penelitian ini juga dinilai dapat digunakan untuk mengatasi *overfitting* dan tetap mempertahankan akurasi yang tinggi.

Saran yang dapat diberikan untuk penelitian selanjutnya berdasarkan penelitian yang sudah dilakukan ini adalah.

- Berdasarkan penelitian ini yang menggunakan kombinasi beberapa *hyper parameter*, maka penelitian selanjutnya dapat dilakukan dengan menambahkan kombinasi *tuning hyper parameter* sehingga dapat memaksimalkan akurasi dari model yang dibuat.
- *Dataset* yang digunakan dalam penelitian selanjutnya dapat dilakukan dengan *dataset* lain, ataupun menggunakan gabungan *dataset* seperti gabungan dari *dataset cleveland* dan *statelog*, sehingga data yang digunakan semakin banyak dan model yang dihasilkan dapat memiliki akurasi yang lebih baik lagi.

Daftar Pustaka

[1] L. Marleni and A. Alhabib, "Faktor Risiko Penyakit Jantung Koroner di RSI SITI Khadijah Palembang," *J. Kesehat.*, vol. 8, no. 3, p. 478, 2017, doi: 10.26630/jk.v8i3.663.

[2] A. Septadaya, C. Dewi, and B. Rahayudi, "Implementasi Extreme Learning Machine dan Fast Independent Component Analysis untuk Klasifikasi Aritmia Berdasarkan Rekaman Elektrokardiogram," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. e-ISSN*, vol. 2548, no. 5, p. 964X, 2019.

[3] E. Prasetyo and B. Prasetiyo, "Peningkatan Akurasi Klasifikasi Algoritma C4.5 Menggunakan Teknik Bagging Pada Diagnosis Penyakit Jantung," vol. 7, no. 5, pp. 1035–1040, 2020, doi: 10.25126/jtiik.202072379.

[4] P. D. Putra and D. P. Rini, "Prediksi Penyakit Jantung dengan Algoritma Klasifikasi," *Pros. Annu. Res. Semin. 2019*, vol. 5, no. 1, pp. 978–979, 2019.

[5] 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, doi: 10.35760/tr.2019.v24i3.2393.

[6] Z. A. Haqie, R. E. Nadiah, and O. P. Ariyani, "Inovasi Pelayanan Publik Suroboyo Bis Di Kota Surabaya," *JPSI (Journal Public Sect. Innov.*, vol. 5, no. 1, p. 23, 2020, doi: 10.26740/jpsi.v5n1.p23-30.

[7] M. S. Dr. Bernatal Saragih, S.P., *Kolesterol dan Usaha-Usaha Penurunannya*, 1st ed.,

no. September. Yogyakarta: Penerbit Bimotry Yogyakarta, 2011.

[8] Rosmalinda, D. Karim, and A. P. Dewi, "Gambaran Tingkat Pengetahuan Perawat Irna Medikal Dalam Menginterpretasi Hasil EKG," no. 1, 2014.

[9] N. Nahdliyah, "Penelitian Tentang Detak Jantung," *Jur. Sist. Komput. Univ. Sriwij.*, vol. 52, no. 1, pp. 1–5, 2019.

[10] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, no. November 2018, 2019, doi: 10.1016/j.imu.2019.100203.

[11] O. W. Purbo and P. Sudiarta, "Inovasi Teknologi Informasi dan Komunikasi Dalam Menunjang Technopreneurship," *Angew. Chemie Int. Ed. 6(11), 951–952.*, pp. 5–24, 2015.

[12] I. D. G. H. Wisana, "Identifikasi Isyarat Elektrokardiogram Segmen ST dan Kontraksi Ventrikel Prematur Berbasis Gelombang Singkat," *Univ. Gadjah Mada Yogyakarta*, 2013.

[13] M. S. K. Ayu, "Proteksi Radiasi Pada Pasien, Pekerja, dan Lingkungan di Dalam Instalasi Radiologi," *Inst. Ilmu Kesehat. Str. Indones.*, 2019.

[14] W. A. Mustofa, "Manfaat Foto Rongten dan Dampaknya," *Intitut Ilmu Kesehat. Str. Indones.*, 2021.

[15] Wahyu Kusuma, "Self Acceptance Pada Remaja Penderita Thalasemia," pp. 8–27, 2016.

[16] T. T. Hanifa, S. Al-faraby, and Adiwijaya, "Analisis Churn Prediction pada Data Pelanggan PT . Telekomunikasi dengan Logistic Regression dan Underbagging," *Univ. Telkom*, vol. 4, no. 2, p. 78, 2017.