KOMPARASI MODEL DECISION TREE DAN RANDOM FOREST UNTUK MEMPREDIKSI PENYAKIT JANTUNG

[1]Mia
Universitas Buana Perjuangan Karawang
Karawang, Indonesia, 41316
if19.mia@mhs.ubpkarawang.ac.id &
085697611617

[2]Anis Fitri Nur Masruriyah
Universitas Buana Perjuangan Karawang
Karawang, Indonesia, 41316
anis.masruriyah@ubpkarawang.ac.id &
085646000302

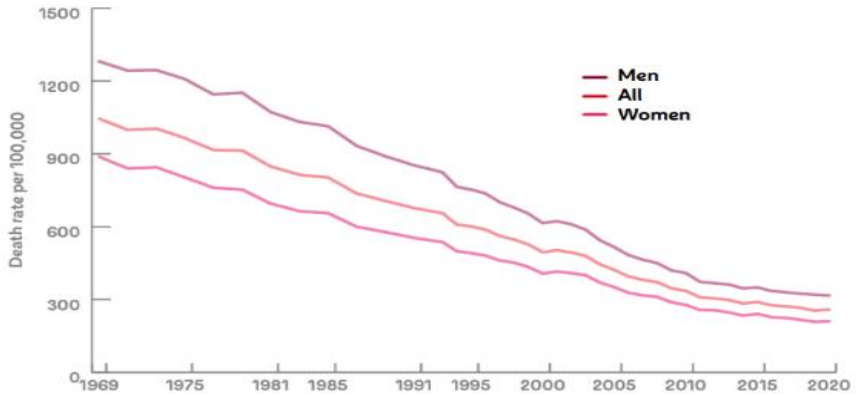
[3]Adi Rizky Pratama
Universitas Buana Perjuangan Karawang
Karawang, Indonesia, 41316
adi.rizky@ubpkarawang.ac.id &
089699971950

Abstract*— Data pasien penyakit jantung yang diperoleh dari Kementrian Kesehatan Republik Indonesia tahun 2020 mejelaskan bahwa, penyakit jantung mengalami peningkatan setiap tahunnya dan menempati peringkat tertinggi penyebab kematian di Indonesia terutama pada usia-usia produktif. Apabila penderita penyakit jantung tidak ditangani dengan baik, maka di usia produktif seorang pasien bisa mengalami kematian lebih cepat. Sehingga, perlunya sebuah model prediksi yang mampu membantu tenaga medis untuk menyelesaikan masalah-masalah Kesehatan. Menggunakan proses klasifikasi algoritma *Random Forest* dan *Decision Tree* dengan mengola data tersebut. Tujuan penelitian untuk mengetahui performa teknik model algoritma machine learning pada algoritma *Decision Tree* C45 dan *Random Forest* Clasifier. Pengunaan teknik confusion matrik untuk pengujian Precision, Recall, dan F1-SCORE, serta *Accuracy*. Berdasarkan hasil penelitian yang telah dilakukan teknik pengujian *K-Fold* 10 dengn teknik Cofusion matrix, *Random Forest* merupakan salah satu teknik prediksi terbaik yang memiliki Accurasy lebih besar 90,72% dan evaluasi menggunakan *Receiver Operating Characteristics* (ROC) curve untuk mengetahui nilai kinerja suatu algoritma dengan nilai Area Under Curve (AUC) pada model sebesar 0,801. Dibandingkan penerapan model prediksi mengguanakan algoritma *Decision Tree* C45 tingkat prediksi dengan *accuracy* sebesar 86,5 dan nilai kinerja algoritma dengan hasil *Area Under Curve* (AUC) yang kurang baik sebesar 0,588. Berdasarkan evaluasi penelitian yang telah dilakukan pada data penyakit jantung, algoritma *Random Forest* sangat cocok untuk prediksi data penyakit jantung yang berasal dari *Centers for Disease Control and Prevention* (2020) yang mampu menghasilkan model prediksi yang lebih baik dengan teknik confucion matrix serta perhitungan *K-Fold cross validation.

Kata kunci — ***Random Forest, K-Fold Cross Validation, Confusion Matrix,* Penyakit Jantung**

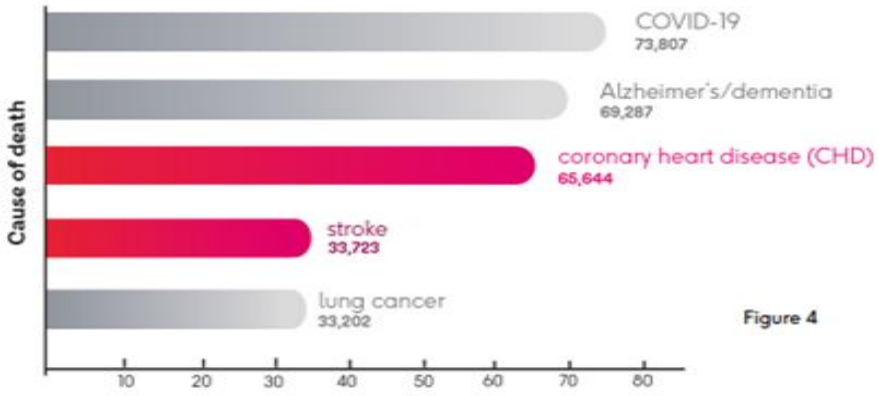
I. PENDAHULUAN (*HEADING 1*)

Berdasarkan kumpulan data yang diperoleh dari Kementrian Kesehatan Republik Indonesia[1]. Penyakit jantung mengalami peningkatan setiap tahunnya dan menempati peringkat tertinggi penyebab kematian di Indonesia terutama pada usia-usia produktif. Selama pandemi COVID-19, pasien dengan penyakit jantung bawaan memiliki risiko keselamatan yang lebih besar karena dapat menyebabkan eksaserbasi dan bahkan kematian. Apabila penderita penyakit jantung tidak ditangani dengan baik, maka di usia produktif seorang pasien bisa mengalami kematian lebih cepat. Berdasarkan hasil penelitian yang telah dilakukan oleh *British Heart Foundation* (BHF) [2] terlampir pada Gambar 1 grafik menjelaskan tentang kematian dini akibat penyakit jantung dan peredaran darah (sebelum usia 75) paling umum di utara Inggris. Tingkat kematian mengambil struktur usia (demografi) daerah setempat diperhitungkan untuk mengungkapkan perbedaan nyata dalam statistik**.**



Gambar 1. *Heart Disease Chart 1969-2020*

Selanjutnya, Penyakit jantung merupakan jenis penyakit jantung dan peredaran darah yang paling umum. Ini terjadi ketika arteri coroner menjadi menyempit oleh penumpukan ateroma, bahan berlemak di dalam dindingnya[2]. Terlampir Pada Gambar 2 grafik penyakit jantung Koroner tahun 2020 cukup besar sekitar 65,644%.



Gambar 2. *Coronary Heart Disease Chart 2020*

Sehingga, perlunya sebuah model prediksi yang mampu membantu tenaga medis untuk menyelesaikan masalah-masalah Kesehatan. Pada penelitian sebelumnya yang dilakukan oleh Pangaribuan[3] membandingkan metode algoritma C45 dan extreme machine learning yang dapat memberikan hasil diagnostik penyakit jantung yang sangat baik hinggga 99.05%. Di sisi lain, penelitian sebelumnya yang dilakukan oleh Rohman dan Rochcham[4] model prediksi penyakit jantung yang dilakukan menggunakan algoritma *Decision Tree* C45 mendapatkan nilai 86,59 %, nilai AUC yang diperoleh 0.957 dan termasuk kategori kelompok sangat baik. Adapun, pada penelitian kanker kulit proses yang digunakan untuk extrasi fiture antara lain: *histogram*, *haralick* dan *hue moments*. Dari ketiga ekstraksi fitur yang digunakan, hasil akurasi algoritma *Random Forest* terbaik diperoleh dengan ekstraksi fitur *hue moments* dengan nilai akurasi 0,842[5]. Selanjutnya, penelitian yang dilakukan oleh Ath[6] untuk memprediksi penyakit jantung menggunakan algoritma *Machine Learning* (ML) sebagai langkah preventif dini pada sistem informasi berbasis desktop. Dengan nilai akurasi yang didapatkan menggunakan metode *Random Forest* dan *Logistic Regression* sebesar 84,48% yang meningkat sebesar 1,32%. Kemudian, penelitian yang dilakukan oleh El-Hasnony[7] membuat model untuk pencegahan penyakit stroke dan jantung menggunakan algoritma *machine*. *Active learning* diterapkan pada penelitian untuk menentukan faktor yang paling berpengaruh pada penyakit jantung.

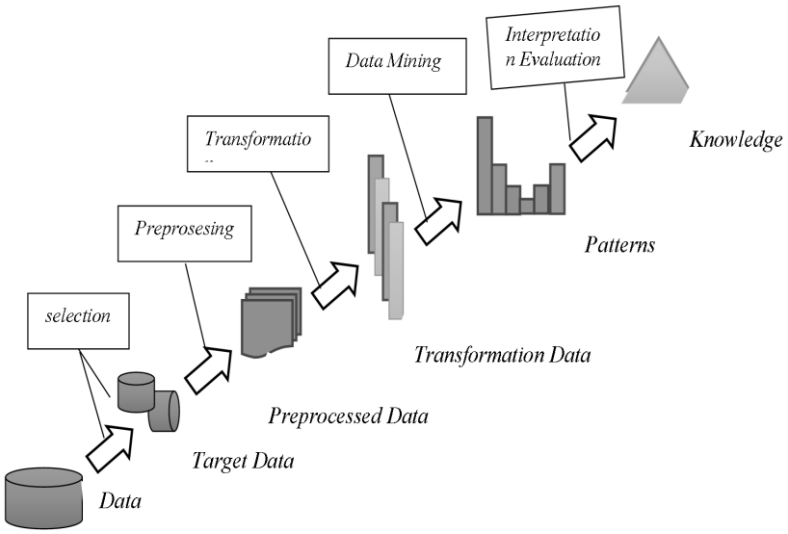
Data yang disimpan dapat digunakan sebagai sumber untuk memprediksi kemungkinan penyakit di masa depan yang membuat teknik penambangan data memainkan peran sentral untuk ekstraksi pengetahuan dan prediksi. Sehingga, perlunya pembuktian model prediksi yang mampu membantu tenaga medis untuk menyelesaikan masalah-masalah kesehatan. Dalam studi sebelumnya, para peneliti menyatakan upaya mereka untuk menemukan model prediksi terbaik, dengan mengusulkan sistem prediksi penyakit jantung[8]. Berdasarkan faktor-faktor yang mempengaruhi tersebut, tenaga medis dapat mengambil langkah-langkah yang tepat untuk mencegah penyakit jantung. Selain itu, penggunaan ekstraksi fitur untuk menentukan variabel yang paling mempengaruhi penyakit jantung berdasarkan perhitungan

II. METODE PENELITIAN

2.1 Alur Penelitian

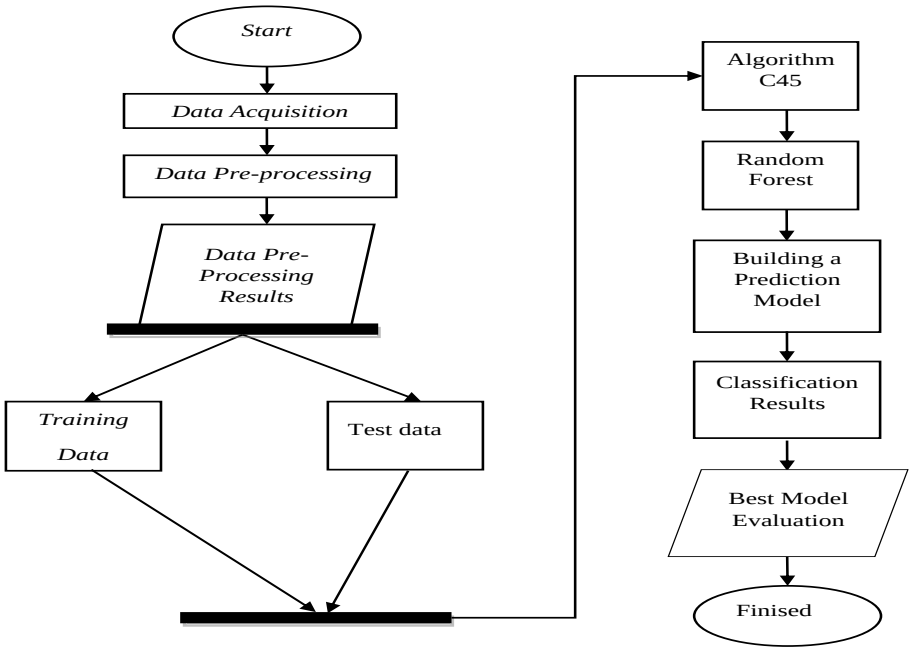
Waktu pencarian dan pengolahan data Penyakit Jantung dilakukan selama 4 Juli hingga 6 Agustus 2022. Selama melakukan penelitian dan analisis data dilakukan diruangan Lab lantai 1 Di Universitas Buana Perjuangan Karawang. Data pasien penyakit jantung dengan total objek lebih dari 300.000 data dengan tujuh variabel dan satu kelas target. Penelitian tugas akhir ini jenis data yang digunakan yaitu data kuantitatif. Sedangkan kumpulan data didapatkan terdiri dari 18 atribut dari catatan medis yang telah dietujui oleh Organisasi Kesehatan Dunia dan dapat diakses pada *Centers for Disease Control and Prevention* (CDC)[9]. Atribut yang dimiliki terdiri dari (*BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer, HeartDisease).* Data yang didapat sesuai dengan proses data mining.

Data Mining adalah proses menemukan pola atau informasi yang menarik dalam data yang dipilih dengan menggunakan teknik atau metode tertentu. Metode, teknik, dan algoritma pada Data Mining sangat bervariasi. Mengumpulkan data, mengambil data, dan menganalisis data mengacu pada mengekstraksi sampel dari kumpulan data menggunakan teknik Data Mining. Populasi yang besar untuk membuat inferensi statistik digunakan untuk memvalidasi pola yang ditemukan[10]. Data Mining memiliki beberapa Teknik yang digunakan untuk memandu seluruh proses yang dilakukan untuk menggabungkan teknik dan teknik dari berbagai disiplin ilmu seperti statistik, *database*, pembelajaran mesin, dan visualisasi. Selanjutnya lampir alur Data Mining pada Gambar 1. yang digunakan untuk memandu seluruh proses.



Gambar 3. *Proses Alur Data Mining*
*(**Sumber**:Fahlevi [11])*

Secara umum, proses analisis data dimulai dengan preprocessing dan ekstraksi fitur hingga dihasilkan pengetahuan baru. Penelitian ini menggunakan empat tahap analitika (*data quality analytics, descpritive analytics, diagnostic analytics* dan *predictive analytics*). Pada tahap pertama pra-pemrosesan data sudah termasuk data *quality analytics* dan *descpritive analytics*. Selanjutnya hasil pra-pemrosesan data diolah untuk mendapatkan hasil *diagnostic analytics* dan *predictive analytics*. Teknik klasifikasi menggunakan *Random Forest* dan *Decision Tree* C45, Berikut ditampilkan *flowchart* aliran pada Gambar 3, fokus penelitian lebih pada pemecahan masalah dan mencapai tujuan penelitian.



Gambar 4. *Flowchart Diagram Aliran*

Dalam proses pengujiannya akan dibandingkan penerapan Random Forest, dan Decision Tree C45. Evaluasi kinerja pada penelitian ini menggunakan teknik *K-Fold Cross Validation* dan *Confusion Matrix.* Cara kerja dari Teknik *K-Fold Cross Validation* adalah dengan membagi data menjadi data uji dan data latih sebanyak K.

2.2 Algoritma Decission Tree

Algoritma C4.5 merupakan pengembangan dari algoritma pohon keputusan ID3 yang diusulkan oleh Quinlan pada tahun 1983[12]. Cara kerja algorima ini dimulai dengan menilai bobot setiap atribut dengan perhitungan entropi (Persamaan 1), kemudian menghitung keterkaitan antar atribut menggunakan *information gain* (Persamaan 2,3)[13][14]. Selanjutnya, atribut yang memiliki keterkaitan paling tinggi terhadap atribut lainnya akan berfungsi sebagai akar pada pohon keputusan dan atribut lain yang memiliki nilai gain lebih rendah akan menjadi dahan atau daun. Cara ini dilakukan dengan memanfaatkan algoritma *split* (Persamaan 4) agar

atribut yang memaksimalkan *rasio* perolehan informasi dipilih sebagai fitur pemisahan terbaik. Kemudian, jika ukuran pohon terlalu luas, maka ukuran pohon akan dipangkas dengan menggunakan teknik pruning. Secara sederhana, pemangkasan dilakukan berdasarkan nilai terendah keterkaitan antar atribut secara rekursif. Sehingga tidak akan muncul aturan yang sama lebih dari satu pada pohon keputusan yang dibangun. *Pruning* dimulai dengan menyimpulkan pohon keputusan dari data latih, kemudian membangun pohon keputusan hingga data pelatihan *fit* sebaik mungkin dan memungkinkan terjadinya *overfitting*. Selanjutnya, mengbah pohon yang dipelajari menjadi seperangkat aturan yang setara dengan membuat satu aturan untuk setiap jalur dari simpul akar ke simpul daun. Kemudian, memangkas setiap aturan dengan menghapus prasyarat apa pun yang menghasilkan peningkatan akurasi perkiraannya[15]. Terakhir, mengurutkan aturan yang dipangkas berdasarkan perkiraan akurasinya, dan pertimbangkan dalam urutan Tabel 2.1 ini saat mengklasifikasikan instance berikutnya.

Tabel 1. *Tahapan Decision Tree*

	RUMUS DECISION TREE
(1)	$Entropy(s) = \sum_{i=1}^{e} -p_i log_2 p_i$
(2)	$Information_{Atribute}(D) = \sum_{j=1}^{v} \lvert \frac{D_i}{D} \rvert x \text{ Info}(D_i)$
(3)	$Information\ Gain\ (Atribute) = info(D) - Info(D_i)$
(4)	$Splitln\ o_{Atribute}(D) = \sum_{j=1}^{v} \left\lvert \frac{D_i}{D} \right\rvert x\ log_2 \lvert \frac{D_i}{D} \rvert$

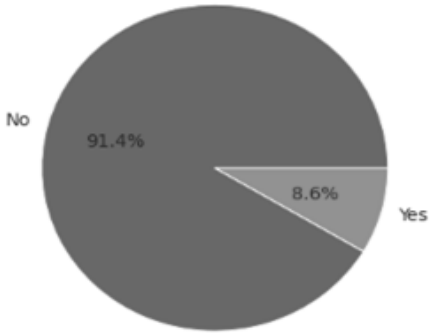
2.3 Algoritma Random Forest

Random Forest classifier merupakan salah satu metode yang digunakan untuk klasifikasi dan regresi[16]. *Random Forest* juga bisa diartikan terbentuk dari sekumpulan *Decision Tree* atau pohon keputusan. Dapat digunakan untuk membuat prediksi kategori menggunakan beberapa kemungkinan nilai dan probabilitas keluaran dapat disesuaikan. Satu hal yang perlu diwaspadai adalah *overfitting*. *Random Forest* bisa terjadi *overfitting*, terutama ketika bekerja dengan kumpulan data yang relatif kecil. Keuntungan menggunakan *Random Forest* dapat mengklasifikasikan data dengan atribut yang tidak lengkap. Digunakan untuk klasifikasi, tetapi tidak terlalu cocok untuk regresi, lebih cocok untuk mengklasifikasi data, dan menangani data sampel yang besar. Pada penelitian seblumnya dilakukan ekstraksi fitur warna menggunakan algoritma *Random Forest*, nilai akurasi terbaik diperoleh dari proses ekstraksi dengan hasil 85% [5].

III. HASIL DAN PEMABAHASAN

3.1 Pre-Processing Data

Hasil dari tahapan prapemrosesan data pada penelitian ini dilakukan dengan cara menghapus data dengan komponen yang tidak lengkap untuk menghindari manipulasi pengisian data lebih jauh karena data yang digunakan lebih dari 1000. Hal ini dilakukan agar data lebih ideal untuk digunakan pada tahap selanjutnya, Selanjutnya setelah data dengan komponen tidak lengkap telah dihapus maka dilakukan normalisasi pada data-data yang memiliki lebih dari 4 kategori. Hal ini dilakukan untuk mengurangi perulangan dan memetakan kejadian yang serupa. Dapat dilihat pada diagram *pie* Gambar 4 menjelaskan bahwa data penyakit jantung tahun 2020 [9] tidak seimbang.



Gambar 5. *Diagram Pie Penyakit Jantung*

Oleh karena itu, sebelum membentuk aturan klasifikasi dan pemodelan, perlu dilakukan pembagian data menjadi dua kelompok yaitu data *train* (*train* model) dan data *test* (menguji performa dari model tersebut). Berbagi data ini bertujuan untuk menganalisis

apakah aturan klasifikasi yang dihasilkan oleh algoritma *Random Forest* dan algoritma *Decision Tree* C45, yang dapat digunakan untuk memprediksi *Heart Desease*.

3.2 Teknik Pengujian

Proses evaluasi pada penelitian ini dengan menggunakan metode *Confusion Matrix* untuk evaluasi mengukur kinerja dalam bentuk persamaan 5 hingga 8. Terletak pada Tabel 3.3 yang dimana Accurasy (Persamaan 5) berisi dari tabel yang terdiri evaluasi dari proses pengujian data untuk melihat kesesuaian dengan menghitung *accuracy*, *Precision* (Persamaan 6) presentase dari label data dengan label positif yang diberikan oleh klasifikasi, *Recall* (Persamaan 7) menghitung *test*ing data set yang benar- benar diprediksi positif oleh model, *F1 Score* (Persamaan 8) menentukan nilai nilai rata – rata *harmonic* dari precision dan recall. Pada tahapan inilah hasil dari proses klasifikasi akan dilihat tingkat kebenarannya. Setelah perhitungan klasifikasi selesai dilakukan selanjutnya akan dievaluasi Confusion Matrix untuk mengetahui performa akurasi, presisi dan recall, dengan rumus sebsgsi berikut.

Tabel 2. *Rumus Confusion Matrix*

	PREDIKSI	
AKTUAL	**A**	**B**
B	**TP_A**	**E_{BA}**
A	**E_{AB}**	**TP_B**

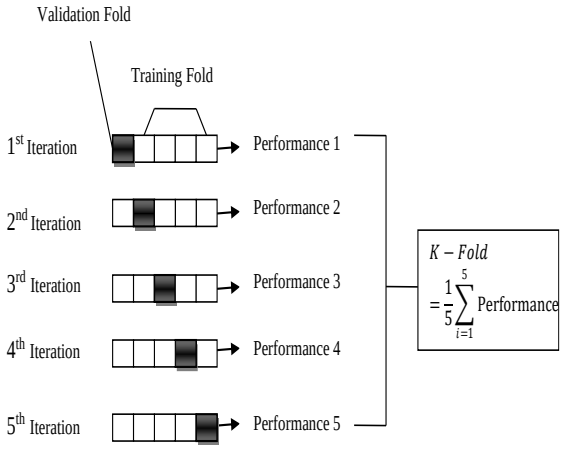
$$accuracy = \frac{True\ Positif + True\ Negatif}{SUM\ The\ Number\ of\ Data}\ x\ 100\% \quad \textbf{(5)}$$

$$Precision = \frac{True\ Posistif}{True\ Positive + False\ Positive}\ x\ 100\% \quad \textbf{(6)}$$

$$recall = \frac{True\ Positif}{True\ Positive + False\ Negatif}\ x\ 100\% \quad \textbf{(7)}$$

$$F1 = 2\ x\ \frac{precission\ x\ recall}{precission + recall} \quad \textbf{(8)}$$

Selanjutnya, pada tahapan ini menggunakan teknik pengujian atau perhitungan *K-Fold* terbaik menggunakan *K-Fold* 10. Pada mechine learning (ML) akan dilakukan split data *train*ing dan teting, untuk mengoptimalkannya menggunakan *Cross Validation* (*K-Fold*). Sebuah metode prosedur pengambilan sampling untuk mendapatkan eveluasi model ML dengan tujuan *Best Validation* dan *Best Learning Result*. Terlampir Ilustrasi pada Gambar 5, dimana model prediksi dimulai dengan membagi seluruh data menjadi data latih dan data uji dengan *K-Fold Cross Validation*, serta dilakukan pengujian secara silang dari masing-masing algoritma. Berikut, tahapan penerapan *K-Fold* yang terlampir pada Gambar 5. Pertama, acak data dan split data menjadi. Kedua, pada iterasi 1 ambil sebagian lekukan menjadi data *train* ing dan sebagian menjadi data *test*ing, lakukan sampai semua lekukan *K-Fold* telah dilakukan. Proses tahapan kedua tersebut dilakukan sebanyak K kali hingga setiap kelompok dilakukan sebagai validasi dan tersisa sebagai data latih. Ketiga, hitung akurasi pada masing – masing lekukan. Keempat, rata – ratakan akurasi setiap iterasi dan menghasilkan acurasy terbaik.



Gambar 6. ***K-Fold Ilustration***
(Sumber :Zitao (2020) [17])

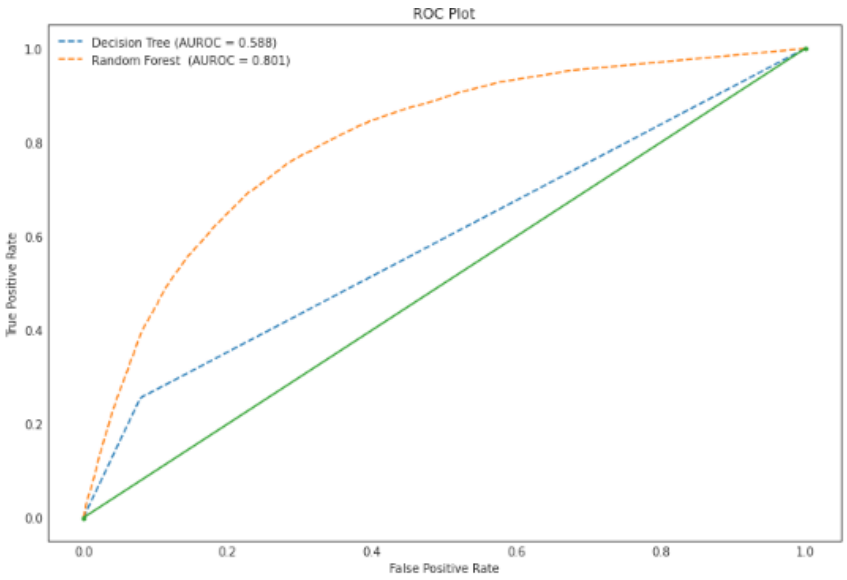
3.3 Evaluasi

Evaluasi didasarkan pada akurasi, presisi, sensitivitas, dan spesifisitas. Proses evaluasi menggunakan data uji yang telah dipisahkan pada proses sebelumnya dan hasil evaluasi menggunakan matriks konfusi yang ditunjukkan pada Table.

Tabel 3. *Hasil Accuracy*

Algorithm	***Accuracy K-Fold* (%)**
Decision Tree C45	86,74
RANDOM FOREST	90,56

Model yang dihasilkan dengan menjalankan Persamaan 1 hingga Persamaan 4 menggunakan *Confusion Matrix* dan perhitungan *K-Fold Cross Validation* 10 pada algoritma *Decision Tree* C45 yang menghasilkan *accuracy* 86,74%. Di sisi lain, algoritma *Random Forest* berhasil membangun model prediksi dengan nilai akurasi lebih tinggi dibanding dengan C4.5 yaitu sebesar 90,56%. Sehingga, berdasarkan hasil perhitungan *Random Forest* (RF) lebih baik dari pada *Decision Tree* C45 untuk memprediksi data penyakit jantung. ROC *Area* menunjukkan metrik seberapa baik model dalam memprediksi. ROC (*Receiver Operating Characteristics*) yang dapat digunakan untuk menganalisis kinerja pada suatu model algoritma [16].



Gambar 7. Hasil *Receiver Operating Characteristics* (ROC)

Dari hasil pengujian di atas, evaluasi menggunakan *counfusion matrix* dan ROC curve untuk mengetahui nilai kinerja suatu model bersarkan nilain AUCnya terbukti bahwa hasil pengujian algoritma *Random Forest* memiliki nilai akurasi yang lebih tinggi jika dibandingkan dengan hasil prediksi algoritma *Decision Tree*. Pada tampilan di atas kinerja model *Random Forest* Jauh lebih baik bersarkan hasil nilai Area Under the Curve (AUC) sebesar 80,1%.

Tabel 4. Perbandingan Accurasy Terbaik

K-Folad Cross Validation		
Algorithm	**Accurasy**	**AUC**
Random Forest	90,72	80,1
Decision Tree C45	86,55	58,8

Nilai akurasi untuk model algoritma C4.5 accurasy sebesar 86,55% dan hasil kinerja model pada algoritma C4.5 berbasis ROC sebesar 58,8%, sedangkan *Random Forest* memiliki akurasi signifikan dari *Decision Tree* sebesar 90,72%. Dengan selisih akurasi 0,122, dapat dilihat pada Tabel diatas. Untuk evaluasi menggunakan *Receiver Operating Characteristics* (ROC) curve untuk mengetahui kinerja suatu model menghasilkan nilai AUC (Area Under Curve) untuk model algoritma *Random Forest* nilai 0,801 dan algoritma *Decision Tree* menghasilkan nilai 0,588 dengan nilai diagnosa Excellent Classification[18] dan selisih nilai keduanya sebesar 0.213. Jadi akurasi terbaik untuk memprediksi penyakit jantung adalah Algoritma *Random Forest* memiliki *accuracy* sebesar 90,72% dan Kurva ROC dengan nilai Area Under the Curve (AUC) 80,1%.

IV. KESIMPULAN

Berdasarkan kesimpulan dari hasil penelitian yang telah dilakukan sebelumnya, algoritma *Random Forest* Clasifier menjadi model perbandingan algoritma terbaik menggunakan perhitungan Confusion Matrix dengan hasil acurassy 90,72%. Penerapan algoritma *Random Forest* Clasifier menggunakan *K-Fold* 10 *Cross Validation* merupakan teknik perhitungan terbaik dalam penerapan model algoritma *Random Forest* dengan hasil 90,72% dan penerapan model algoritma *Decision Tree* 86,55%. Berdasarkan teknik pengujian *K-Fold* 10 *Random Forest* merupakan salah satu teknik perhitungan terbaik yang memiliki Accurasy lebih besar 90,72% dibandingkan penerapan *Decision Tree* dalam memprediksi data penyakit jantung. Pada pengelompokan akurasi data mining maka akurasi 0.90-1.00 = *Excellent* classification. Dengan hasil *Receiver Operating Characteristics* (ROC) untuk mengukur suatu nilai kinerja model jauh lebih efisien dari *Decision Tree* dengan nili Area Under the Curve (AUC) sebesar 80,1% menngunakan algoritma Random Forest. Berdasarkan hasil dan analisis terkait implementasi random forest, maka bisa disimpulkan bahwa *Random Forest* mampu digunakan untuk menghasilkan model prediksi penyakit jantung. Penggunaan teknik penentuan *train* ing data dan *test* ing data menggunakan kesuluruhan data mampu menghasilkan model yang lebih baik dengan teknik pembagian data *K-Fold cross validation.*

DAFTAR PUSTAKA

[1] "Kementerian Kesehatan Republik Indonesia." https://www.kemkes.go.id/index.php (accessed Aug. 06, 2022).

[2] BHF, "UK Factsheet," *Br. Hear. Found.*, no. April, pp. 1–21, 2019.

[3] J. J. Pangaribuan, C. Tedja, and S. Wibowo, "PERBANDINGAN METODE ALGORITMA C4.5 DAN EXTREME LEARNING MACHINE UNTUK MENDIAGNOSIS PENYAKIT JANTUNG KORONER," 2019.

[4] A. Rohman and D. M. Rochcham, "MODEL ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT JANTUNG," 2018.

[5] N. Khasanah, R. Komarudin, N. Afni, Y. I. Maulana, and A. Salim, "Skin Cancer Classification Using *Random Forest* Algorithm," *Sisfotenika*, vol. 11, no. 2, p. 137, 2021, doi: 10.30700/jst.v11i2.1122.

[6] S. Ath *et al.*, "Jurnal Teknologi Terpadu HYBRID MACHINE LEARNING MODEL UNTUK MEMPREDIKSI PENYAKIT JANTUNG DENGAN METODE LOGISTIC REGRESSION DAN RANDOM," vol. 8, no. 1, pp. 40–46, 2022.

[7] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction," *Sensors*, vol. 22, no. 3, 2022, doi: 10.3390/s22031184.

[8] D. Derisma, "Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining," *J. Appl. Informatics Comput.*, vol. 4, no. 1, pp. 84–88, 2020, doi: 10.30871/jaic.v4i1.2152.

[9] "CDC - 2020 BRFSS Survey Data and Documentation." https://www.cdc.gov/brfss/annual_data/annual_2020.html (accessed Sep. 01, 2022).

[10] S. Polamuri, "HOW THE LOGISTIC REGRESSION MODEL WORKS," *Dataaspirant*, 2017. https://dataaspirant.com/how-logistic-regression-model-works/ (accessed Aug. 04, 2022).

[11] A. Fahlevi, "Proses Data Mining KDD," 2021. https://sis.binus.ac.id/2021/09/30/proses-data-mining-kdd/.

[12] W. Sullivan, "Machine Learning For Beginners Guide Algorithms," 2017. https://www.perlego.com/book/975057/machine-learning-for-beginners-guide-algorithms-supervised-unsupervsied-learning-decision-tree-random-forest-introduction-pdf.

[13] A. Cherfi, K. Nouira, and A. Ferchichi, "Very Fast C4.5 *Decision Tree* Algorithm," *Appl. Artif. Intell.*, vol. 32, no. 2, pp. 119–137, 2018, doi:

10.1080/08839514.2018.1447479.

[14] M. Kretowski, “Evolutionary Decision Trees in Large-Scale Data Mining,” vol. 59, 2019, doi: 10.1007/978-3-030-21851-5.

[15] M. Mia, A. F. N. Masruriyah, and A. R. Pratama, “The Utilization of *Decision Tree* Algorithm In Order to Predict Heart Disease,” *J. Sisfotek Glob.*, vol. 12, no. 2, p. 138, 2022, doi: 10.38101/sisfotek.v12i2.551.

[16] A. Primajaya and B. N. Sari, “*Random Forest* Algorithm for Prediction of Precipitation,” *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[17] S. Zitao, “3 min of Machine Learning: Cross Vaildation,” *Zitao’s Web*, 2020. https://zitaoshen.rbind.io/project/machine_learning/machine-learning-101-cross-vaildation/ (accessed Aug. 31, 2022).

[18] A. Rohman and D. M. Rochcham, “MODEL ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT JANTUNG,” *J. Neo Tek.*, vol. 4, no. 2, pp. 52–54, 2018.