Prediksi Risiko Penyakit Jantung Dengan Algoritma Random Forest

Anak Agung Bagus Jelantik Kusuma[1], Gede Alpina Dendra[2], Achmad Irsyadul Ibad[3]

[1,2,3]Universitas Pendidikan Ganesha; Jalan Udayana No. 11, (0362)22570, Singaraja

[1,2,3]Program Studi Ilmu Komputer, Jurusan Teknik Informatika, Fakultas Teknik dan Kejuruan, Undiksha, Indonesia

e-mail: *[1]anak.agung.bagus@student.undiksha.ac.id, [2]alpina@student.undiksha.ac.id, [3]achmad@student.undiksha.ac.id

Abstrak

Penyakit jantung merupakan salah satu penyebab utama kematian di seluruh dunia. Deteksi dini terhadap risiko penyakit jantung berperan sangat penting bagi kesehatan masyarakat untuk mengurangi dampaknya. Penelitian ini menggunakan pendekatan machine learning untuk memprediksi risiko penyakit jantung berdasarkan data klinis pasien, seperti usia, jenis kelamin, tekanan darah, kolesterol serum, denyut jantung maksimum, dan kadar gula darah. Model yang digunakan adalah Random Forest Classifier, yang memiliki keunggulan dalam menangani data tabular dengan hubungan non-linear dan memberikan hasil prediksi yang akurat. Dataset yang digunakan dalam penelitian ini adalah Cardiovascular Disease Dataset, yang didapatkan dari Kaggle dan mencakup berbagai fitur klinis serta klasifikasi risiko sebagai target. Setelah dilakukan pra pemrosesan, termasuk penanganan nilai hilang, normalisasi data, dan pembagian dataset, model dilatih menggunakan data latih (70%) dan dievaluasi dengan data uji (30%). Hasil evaluasi menunjukkan bahwa model Random Forest mencapai akurasi tinggi, dengan matriks precision, recall, dan F1-score yang konsisten. Analisis feature importance menunjukkan bahwa fitur seperti kolesterol serum dan tekanan darah istirahat memiliki pengaruh signifikan terhadap prediksi risiko. Penelitian ini membuktikan bahwa model Random Forest dapat menjadi alat yang andal untuk membantu prediksi risiko penyakit jantung.

Kata kunci*— Penyakit Jantung, Random Forest, Klasifikasi, Prediksi.*

Abstract

Heart disease is one of the leading causes of death worldwide. Early detection of heart disease risk plays a crucial role in public health to mitigate its impact. This research uses a machine learning approach to predict heart disease risk based on patients' clinical data, such as age, gender, blood pressure, serum cholesterol, maximum heart rate, and blood sugar levels. The model used is the Random Forest Classifier, which excels at handling tabular data with non-linear relationships and provides accurate prediction results. The dataset used in this study is the Cardiovascular Disease Dataset, obtained from Kaggle, which includes various clinical features and risk classification as the target. After preprocessing, including handling missing values, data normalization, and dataset splitting, the model was trained using training data (70%) and evaluated with test data (30%). The evaluation results show that the Random Forest

model achieves high accuracy, with consistent precision, recall, and F1-score metrics. Feature importance analysis shows that features such as serum cholesterol and resting blood pressure have a significant influence on risk prediction. This research proves that the Random Forest model can be a reliable tool to aid in heart disease risk prediction.

Keywords— *Heart Disease, Random Forest, Classification, Prediction.*

1. INTRODUCTION

Penyakit jantung masih menjadi salah satu tantangan terbesar dalam dunia kesehatan, dengan tingkat kematian yang tinggi di seluruh dunia dan dampak signifikan terhadap sistem pelayanan kesehatan [1]. Penyakit ini terjadi ketika jantung gagal berfungsi secara optimal, yang sering kali disebabkan oleh penyempitan arteri koroner akibat faktor risiko seperti hipertensi, kolesterol tinggi, obesitas, dan diabetes [2]. Faktor risiko tersebut dapat dikategorikan menjadi dua jenis, yaitu faktor risiko yang tidak dapat diubah seperti usia dan genetik, serta faktor risiko yang dapat diubah seperti pola makan yang buruk, kurang aktivitas fisik, dan konsumsi alkohol berlebih [3].

Kemajuan dalam teknologi pembelajaran mesin (machine learning) telah memberikan solusi inovatif dalam mendeteksi penyakit jantung secara dini dan meningkatkan akurasi prediksi. Salah satu algoritma yang banyak digunakan adalah Artificial Neural Network (ANN), yang terbukti mampu mengklasifikasikan penyakit jantung dengan akurasi tinggi berdasarkan data klinis dan demografis [2]. Selain itu, algoritma Naive Bayes telah diterapkan secara luas dalam sistem diagnostik medis karena kesederhanaan dan efektivitasnya dalam menangani data yang terstruktur [5]. Algoritma K-Nearest Neighbor (KNN) juga menjadi pilihan populer dalam klasifikasi risiko penyakit jantung, terutama karena kemampuannya dalam menganalisis dataset besar dengan presisi yang baik [6]. Di sisi lain, algoritma Random Forest dikenal karena pendekatan ensemble-nya yang kuat, mampu menghasilkan prediksi yang andal dan mengurangi risiko overfitting pada dataset berukuran besar [10]. Linear Discriminant Analysis (LDA) juga digunakan secara efektif untuk mengurangi dimensi data sekaligus meningkatkan kinerja klasifikasi pada kasus prediksi risiko penyakit jantung [9].

Penelitian ini bertujuan untuk mengeksplorasi penerapan algoritma pembelajaran mesin yang lebih canggih dalam prediksi penyakit jantung, dengan mengacu pada penelitian sebelumnya yang telah menunjukkan efektivitasnya [6]. Dengan mengintegrasikan teknik baru dan dataset yang beragam, diharapkan hasil penelitian ini dapat memberikan kontribusi signifikan dalam pengembangan sistem deteksi dini dan prediksi risiko penyakit jantung, sehingga dapat menurunkan beban kesehatan secara global.

2. METHODS

2.1 Problem Analysis

Penyakit jantung adalah salah satu penyebab utama kematian di seluruh dunia. Deteksi dini risiko penyakit jantung dapat membantu pencegahan dan pengelolaan kondisi pasien dengan lebih efektif. Namun, identifikasi faktor risiko seringkali membutuhkan analisis kompleks terhadap berbagai data klinis, seperti usia, tekanan darah, kadar kolesterol, dan denyut jantung. Pendekatan manual memakan waktu dan rentan terhadap kesalahan manusia.

Model prediksi dapat memprediksi risiko penyakit jantung secara otomatis berdasarkan data klinis pasien. Pada penulisan ini, diterapkan algoritma machine learning untuk memprediksi apakah seseorang berisiko terkena penyakit jantung dengan klasifikasi risiko menjadi iya dan tidak.

2.2 Design Method

Gambar 1. Flowchart Diagram Pembuatan Random Forest Classification Model Secara Umum

Pembuatan model dimulai dengan mencari dan memilih dataset yang relevan. Setelah dataset diperoleh, dilakukan visualisasi data untuk memahami distribusi fitur dan hubungan antar variabel, sehingga mempermudah analisis. Tahap berikutnya adalah data cleaning, termasuk menangani nilai yang hilang dan normalisasi fitur numerik agar lebih seragam. Data kemudian dibagi menjadi data latih, validasi, dan uji untuk memastikan evaluasi model yang adil. Setelah itu, model Random Forest Classifier dibuat dan dilatih menggunakan data latih.

2. 2.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah Cardiovascular Disease Dataset, yang berisi data klinis pasien dengan fitur seperti usia (dalam tahun), jenis kelamin (0 = Perempuan, 1 = Laki-laki), tekanan darah istirahat (dalam mmHg), kolesterol serum (dalam mg/dL), denyut jantung maksimum, kadar gula darah (0 = Normal, 1 = Tinggi > 120 mg/dL). Variabel dependen pada dataset ini adalah outcome. Bernilai 1 jika seseorang beresiko terkena

penyakit jantung dan 0 jika sebaliknya.

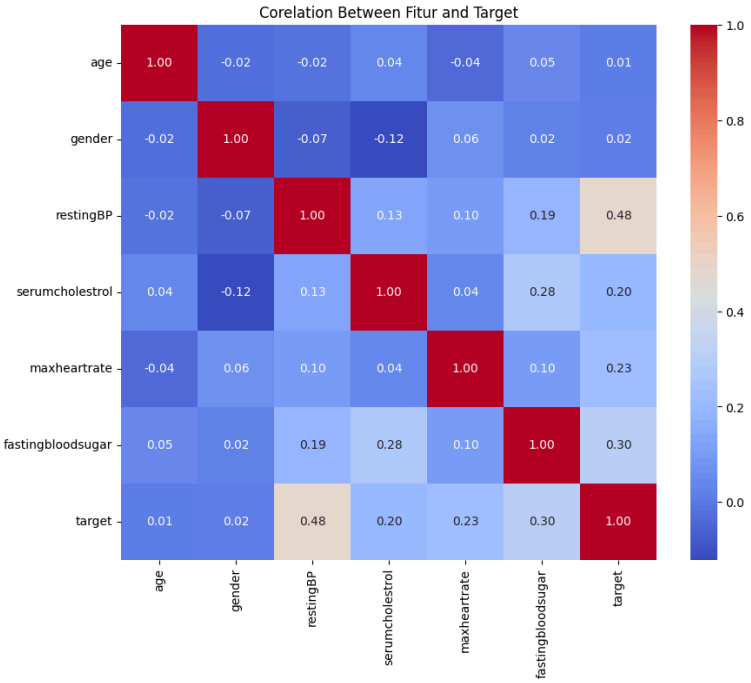
	patientid	age	gender	chestpain	restingBP
count	1.000000e+03	1000.00000	1000.000000	1000.000000	1000.000000
mean	5.048704e+06	49.24200	0.765000	0.980000	151.747000
std	2.895905e+06	17.86473	0.424211	0.953157	29.965228
min	1.033680e+05	20.00000	0.000000	0.000000	94.000000
25%	2.536440e+06	34.00000	1.000000	0.000000	129.000000
50%	4.952508e+06	49.00000	1.000000	1.000000	147.000000
75%	7.681877e+06	64.25000	1.000000	2.000000	181.000000
max	9.990855e+06	80.00000	1.000000	3.000000	200.000000

	serumcholestrol	fastingbloodsugar	restingrelectro	maxheartrate
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	311.447000	0.296000	0.748000	145.477000
std	132.443801	0.456719	0.770123	34.190268
min	0.000000	0.000000	0.000000	71.000000
25%	235.750000	0.000000	0.000000	119.750000
50%	318.000000	0.000000	1.000000	146.000000
75%	404.250000	1.000000	1.000000	175.000000
max	602.000000	1.000000	2.000000	202.000000

	exerciseangia	oldpeak	slope	noofmajorvessels	target
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	0.498000	2.707700	1.540000	1.222000	0.580000
std	0.500246	1.720753	1.003697	0.977585	0.493805
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	1.300000	1.000000	0.000000	0.000000
50%	0.000000	2.400000	2.000000	1.000000	1.000000
75%	1.000000	4.100000	2.000000	2.000000	1.000000
max	1.000000	6.200000	3.000000	3.000000	1.000000

Gambar 2. Deskripsi Cardiovascular Disease Dataset

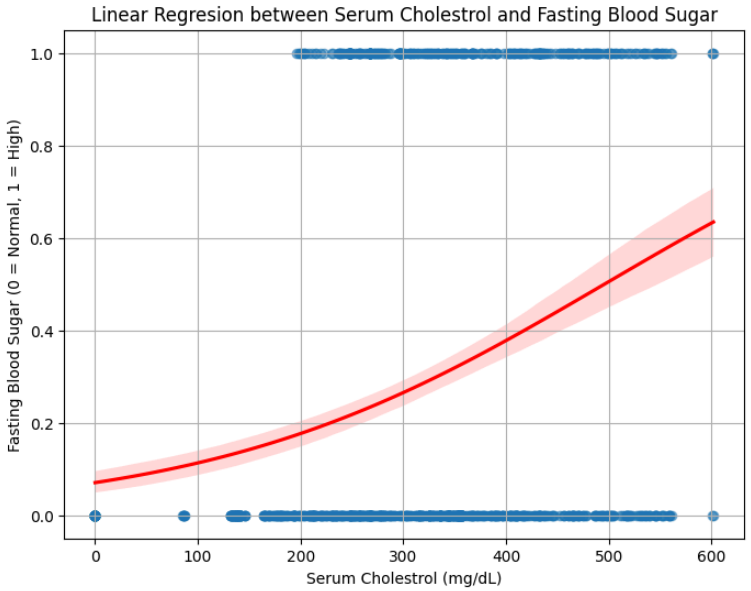
Dataset ini diproses melalui beberapa tahap, seperti pembersihan pada data yang bernilai nol dan normalisasi data yang tidak berdistribusi normal. Visualisasi Heatmap juga digunakan untuk mendapatkan wawasan tentang data dan temuan korelasi antar fitur. Warna merah mewakili lebih banyak korelasi dan warna biru tua mewakili lebih sedikit korelasi.



Gambar 3. Visualisasi Heat Map

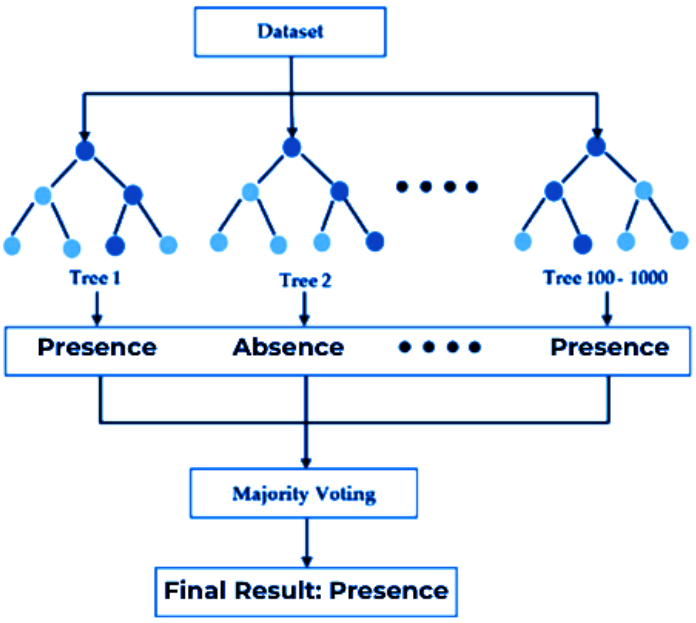
Selanjutnya disosialisasikan korelasi antara serum cholesterol dan fasting blood sugar.

Pada heatmap, dapat dilihat bahwa serum cholesterol dan fasting blood pressure memiliki korelasi yang cukup tinggi yaitu 0.28. Pada Gambar 4 memiliki korelasi linear. Gambar 4 menunjukan bahwa tren kenaikan serum cholesterol akan diikuti dengan kenaikan fasting blood sugar.



Gambar 4. Visualisasi Korelasi Features Serum Cholesterol Dengan Fasting Blood Sugar

2. 2.2 Random Forest Classifier



Gambar 5. Diagram Proses Random Forest Classifier

Salah satu jenis algoritma Random Forest yang digunakan untuk klasifikasi. Algoritma ini bekerja dengan membuat banyak decision tree selama proses pelatihan, lalu menggunakan hasil voting dari pohon-pohon tersebut untuk menentukan kelas akhir dari suatu input data.

Pada dataset yang digunakan, yaitu Cardiovascular Disease Dataset sebanyak 80% data yang ada digunakan untuk training set, dan sebanyak 20% data digunakan untuk testing set. Komputer akan mempelajari dataset dengan mengaplikasikan algoritma random forest classifier.

Bootstrap Sampling adalah metode di mana dataset asli dipecah menjadi beberapa subset menggunakan teknik sampling acak dengan pengembalian. Tiap subset tersebut digunakan untuk melatih satu decision tree. Dalam proses pembuatan decision tree, fitur-fitur dipilih secara acak dalam jumlah tertentu (m) dari total fitur yang tersedia. Decision tree kemudian dibangun berdasarkan subset data dan subset fitur yang dipilih, dengan menggunakan kriteria seperti Gini Index atau Entropy untuk menentukan split terbaik. Setelah semua pohon selesai dilatih, dilakukan proses voting untuk membuat prediksi. Ketika data baru dimasukkan, setiap pohon memberikan prediksinya, dan hasil akhir ditentukan berdasarkan voting mayoritas dari semua pohon, di mana Tk(x) adalah prediksi dari pohon ke-k untuk input x, dan K adalah jumlah total pohon.

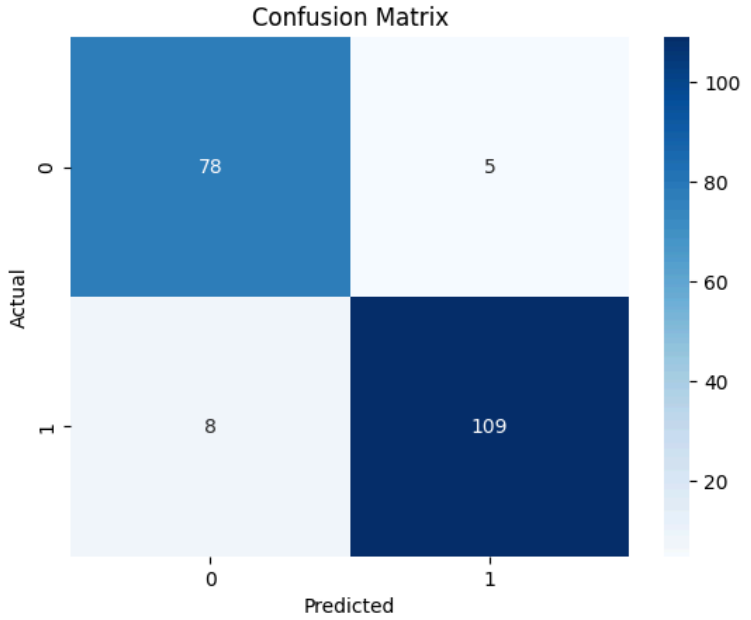
Dalam decision tree, kriteria split seperti Gini Index dihitung menggunakan rumus $Gini = 1 - \sum_{i=1}^{n} p_i^2$ di mana Pi adalah proporsi kelas ke-ii dalam node. Alternatifnya, digunakan Entropy (Information Gain) yang dihitung dengan $Entropy = -\sum_{i=1}^{n} p_i \log_2(p_i)$. Node akan dipisahkan untuk meminimalkan nilai Gini atau Entropy di setiap langkah.

Random Forest memiliki beberapa hyperparameter penting, salah satunya adalah n_estimators, yaitu jumlah pohon dalam hutan. Semakin banyak pohon yang digunakan, semakin akurat prediksi yang dihasilkan, tetapi hal ini juga membutuhkan waktu komputasi yang lebih lama.

3. RESULTS AND DISCUSSION

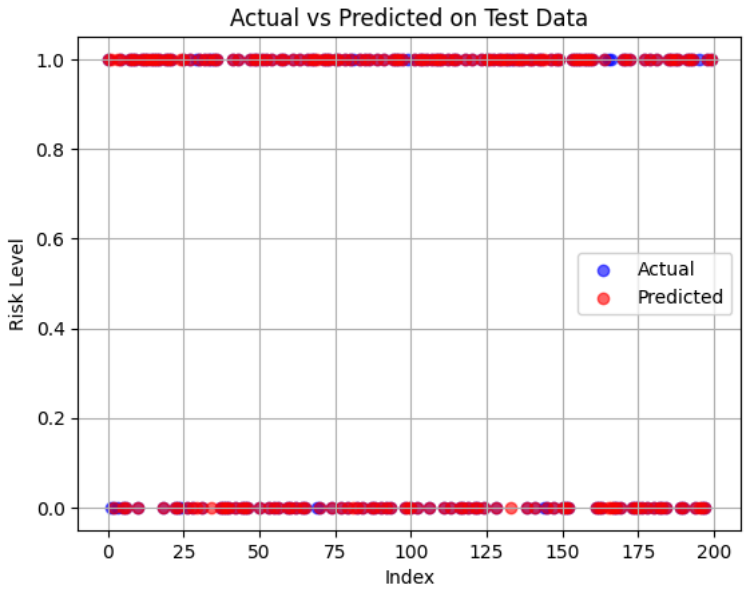
Eksperimen ini dilakukan untuk memprediksi risiko penyakit jantung dengan menggunakan algoritma *random forest classifier.* Dataset yang digunakan terbagi menjadi data training dan data testing, dengan perbandingan data training sebanyak 80% dan data testing sebanyak 20%. Algoritma *random forest classifier* digunakan untuk memprediksi apakah seseorang beresiko terkena penyakit jantung atau tidak dan mengecek akurasi dari prediksi yang dilakukan tersebut.

Pada saat melakukan pengukuran performa model tersebut, terdapat beberapa hal yang perlu diperhatikan : true positive (TP), true negative (TN), false positive (FP) dan false negative (FN). Keempat hal tersebut didapatkan dari confusion matrix pada model yang telah dibuat. Model random forest classifier yang dihasilkan memiliki akurasi sebesar 0.935 atau 93.5% dengan confusion matrix yang ada pada Gambar 6.



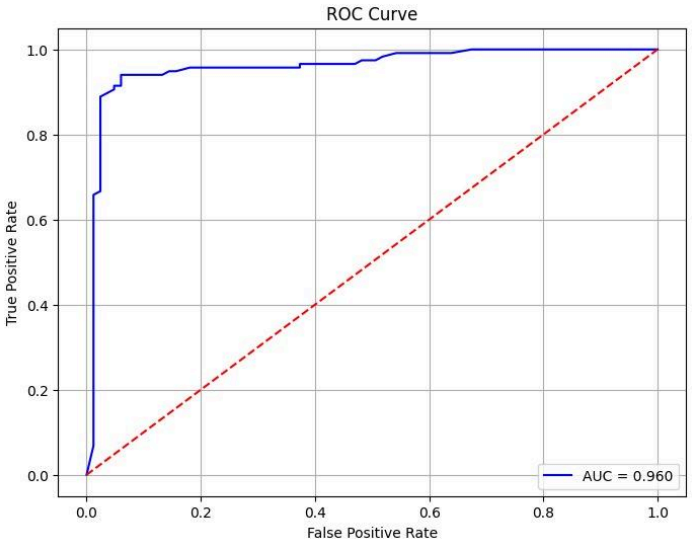
Gambar 6. Confusion Matrix

Selain confusion matrix untuk dapat mengukur performa dari model, dilakukan juga visualisasi scatter plot. Scatter plot memvisualisasikan perbandingan antara nilai aktual dan prediksi pada dataset uji. Hal ini berguna untuk melihat apakah model memprediksi nilai yang mendekati nilai sebenarnya, memahami pola prediksi model terhadap data uji, dan mengidentifikasi perbedaan atau kesalahan prediksi seperti titik-titik merah jauh dari titik-titik biru. Hasilnya akan berupa dua kumpulan titik biru untuk nilai aktual, dan merah untuk prediksi pada grafik, yang memungkinkan kita untuk mengevaluasi kinerja model secara visual.



Gambar 7. Scatter Plot

Untuk dapat mengukur performansi dari model, kita juga dapat menggunakan kurva *Receiver Operating Characteristic* atau *ROC*. kurva *ROC* menggambarkan performansi model klasifikasi tanpa memperhatikan distribusi kelas atau kesalahan, sumbu vertikal pada kurva *ROC* menggambarkan nilai True Positif (TP) dan sumbu horizontal menggambarkan False Positif (FP). kurva *ROC* dari model prediksi diabetes dengan menggunakan logistic regression yang dibuat pada eksperimen ini dapat dilihat pada Gambar 8 *ROC Model Diabetic Logistic Regression.*



Gambar 8. ROC Model Diabetic Logistic Regression.

Performansi klasifikasi model dapat dikatakan baik jika nilai AUC (Area Under Curve) bernilai dalam rentang 0.8 sampai dengan 0.9. Berdasarkan kurva ROC yang telah dihasilkan, dapat dilihat bahwa model yang dirancang menggunakan algoritma Random Forest sudah dapat melakukan klasifikasi dengan sangat baik. Hal ini ditunjukkan oleh nilai AUC yang dihasilkan, yaitu sebesar 0.935, yang berada di atas ambang batas 0.9.

Setelah mengetahui masing-masing nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN), kita dapat menyusun classification report untuk mengevaluasi performansi model yang telah dibuat. Classification report ini terdiri dari beberapa metrik penting: Accuracy, Precision, Recall, F1 Score, dan Support.

Accuracy menggambarkan seberapa akurat model dalam mengklasifikasikan data dengan benar. Akurasi dapat dihitung dengan rumus berikut:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

Precision menggambarkan akurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model. Precision dapat dihitung dengan rumus berikut:

`

$$Precision = \frac{TP}{TP+FP}$$

Recall menggambarkan keberhasilan model dalam menemukan kembali informasi yang relevan. Recall dapat dihitung dengan rumus berikut:

$$Recall = \frac{TP}{TP+FN}$$

F-1 Score menggambarkan perbandingan rata-rata antara Precision dan Recall yang dibobotkan. F1 Score dapat dihitung dengan rumus berikut:

$$F\text{-}1\ Score = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

Tabel 1. *Classification Report*

Diabetic Dataset	*Precision*	*Recall*	*F1-Score*	*Support*
0	0.91	0.94	0.92	83
1	0.96	0.93	0.94	117
Accuracy			0.94	200
Macro Avg	0.93	0.94	0.93	200
Weighted Avg	0.94	0.94	0.94	200

Untuk mengecek performa model, dilakukan uji model dengan memasukkan data pasien baru dan melakukan prediksi untuk mengetahui apakah pasien tersebut diprediksi berisiko mengalami penyakit jantung atau tidak.

Tabel 2. Data Percobaan

Age	*Gender*	*Resting B P*	*Max Heart Rate*	*Fasting Blood Sugar*
45	1	250	160	0

Hasil prediksi model menghasilkan array([1]), yang berarti data pasien pada Tabel Data Percobaan tersebut diprediksi berisiko menderita penyakit jantung.

4. CONCLUSIONS

Penelitian ini berhasil menerapkan metode Random Forest untuk memprediksi risiko serangan jantung dengan akurasi mencapai 93,5%. Model menunjukkan performa yang sangat baik, dengan rata-rata tertimbang Precision, Recall, dan F1-Score sebesar 94%. Faktor utama yang berkontribusi dalam prediksi adalah tekanan darah, kadar kolesterol, dan detak jantung maksimum, sementara usia, gula darah puasa, dan jenis kelamin memberikan pengaruh tambahan.Pada pembuatan model, fitur-fitur yang memiliki korelasi rendah atau tidak signifikan terhadap prediksi seharusnya dieliminasi untuk meningkatkan performa model. Namun, proses tersebut belum dapat dilakukan karena keterbatasan waktu untuk analisis dan evaluasi yang lebih mendalam.Untuk penelitian selanjutnya, diharapkan dapat dilakukan evaluasi yang lebih komprehensif dengan memperhatikan perhitungan nilai error, seperti false positives dan false negatives, yang berasal dari confusion matrix. Hal ini dapat memberikan wawasan tambahan mengenai kesalahan prediksi model, terutama untuk mengidentifikasi kelas risiko yang sulit diprediksi. Selain itu, analisis fitur yang lebih mendalam juga perlu dilakukan untuk

memastikan bahwa hanya fitur signifikan yang digunakan, sehingga dapat meningkatkan efisiensi dan performa model secara keseluruhan.

REFERENCES

[1] Abdul Saboor, Muhammad Usman, Sikandar Ali, Ali Samad, Muhmmad Faisal Abrar, Najeeb Ullah "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms" *Mobile Information Systems* Volume 2022, Article ID 1410169, 9 pages [Online]. Available: https://doi.org/10.1155/2022/1410169. [Accessed: 22-Des-2024].

[2] David Galih Pradana, Muhammad Luthfi Alghifari, Muhammad Farhan Juna, Shulun Dwisiwi Palaguna, "Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network," *Indonesian Journal of Data and Science (IJODAS). Vol 3, No 2, July 2022, pp. 55-60* [Online]. Available: https://doi.org/10.56705/ijodas.v3i2.35. [Accessed: 22-Des-2024].

[3] Sibo Prasad Patro, Gouri Sankar Nayak, Neelamadhab Padhy "Heart disease prediction by using novel optimization algorithm: A supervised learning prospective," *Informatics in Medicine Unlocked*, Volume 26, 2021, 100696 [Online]. Available: https://doi.org/10.1016/j.imu.2021.100696. [Accessed: 22-Des-2024].

[4] Anang Dwi Purwanto, Ketut Wikantika, Albertus Deliar, and Soni Darmawan "Decision Tree and Random Forest Classification Algorithms for Mangrove Forest Mapping in Sembilang National Park, Indonesia," *Remote Sens. 2023, 15, 16.* Available: https://www.mdpi.com/journal/remotesensing. [Accessed: 22-Des-2024].

[5] D. M. A. Hafiz and D. P. Sari, "Sistem Prediksi Penyakit Jantung Menggunakan Metode Naive Bayes," Jurnal Rekayasa Elektro Sriwijaya, vol. 2, no. 2, pp. 152–158, 2021 [Online]. Available: https://jres1.ejournal.unsri.ac.id/index.php/jres/article/download/29/20/23. [Accessed: 15-Dec-2024].

[6] M. Ula and F. Rahmawati, "Implementasi Machine Learning untuk Prediksi Penyakit Jantung Menggunakan Algoritma K-Nearest Neighbor," BIOS: Jurnal Teknologi Informasi dan Rekayasa Komputer, vol. 3, no. 1, pp. 45–52, 2022 [Online]. Available: https://bios.sinergis.org/bios/article/download/152/68/. [Accessed: 17-Dec-2024].

[7] C. B. Sonjaya, A. Fitri, N. Masruriyah, and D. S. Kusumaningrum, "Perbandingan Kinerja Algoritma Klasifikasi Deteksi Penyakit Jantung," INTERNAL (Jurnal Sistem Informasi), vol. 5, no. 2, pp. 166–175, 2022 [Online]. Available: https://jurnalmahasiswa.com/index.php/biikma/article/download/1598/1105/3399 . [Accessed: 19-Dec-2024].

[8] R. Pranandito and H. Harianto, "Perbandingan Prediksi Penyakit Serangan Jantung Menggunakan Model Machine Learning," JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 8, no. 4, pp. 1228–1237, 2023 [Online]. Available: https://www.researchgate.net/publication/377245093. [Accessed: 20-Dec-2024].

[9] R. M. Siahaan and R. Panjaitan, "Prediksi Risiko Penyakit Jantung Menggunakan Algoritma Linear Discriminant Analysis," Jurnal Rekayasa Sistem dan Teknologi Informasi, vol. 1, no. 2, pp. 1–9, 2021 [Online]. Available: https://journal.jisti.unipol.ac.id/index.php/jisti/article/download/222/173. [Accessed: 22-Dec-2024].

[10] I. Suryani and H. R. Rahmat, "Penerapan Algoritma Random Forest untuk Prediksi Risiko Penyakit Jantung Berdasarkan Data Pasien," Jurnal Teknologi Informasi dan Komputer, vol. 7, no. 3, pp. 190–198, 2021 [Online]. Available: https://jurnal.tik.unipol.ac.id/article/view/567. [Accessed: 18-Dec-2024].