Jurnal Minfo Polgan
Volume 12, Nomor 2, November 2023 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13214 p-ISSN : 2089-9424

Terbit : 05 November 2023

Prediksi Penyakit Jantung Menggunakan Algoritma Random Forest

[1]Firmansyah, [2] Agus Yulianto
[1]Universitas Bina Sarana Informatika, [2]Universitas Nusa Mandiri, Indonesia.

[1]firmansyah.fmh@bsi.ac.id, [2]agus.aag@nusamandiri.ac.id

ABSTRAK

Pembelajaran mesin di bidang kesehatan terus berkembang dan banyak digunakan untuk membantu dokter seperti visualisasi, klasifikasi kesehatan, prediksi penyakit dan banyak lagi. Random forest merupakan salah satu algoritma yang digunakan dalam bidang kesehatan yang dapat memprediksi penyakit jantung. Dengan menggunakan random forest, data yang diprediksi lebih akurat dibanding decision tree. Dengan menggunakan metode CRISP-DM, terbukti random forest mampu memprediksi penyakit jantung berdasarkan kebiasaan pasien. Tingkat akurasi yang dimiliki oleh random forest adalah 91% akurat dalam memprediksi penyakit jantung.

Kata Kunci: naïve bayes classifier, klasifikasi naïve bayes, prediksi kelulusan siswa, prediksi.

PENDAHULUAN

Pembelajaran mesin dalam bidang kesehatan terus berkembang hingga saat ini dengan tujuan untuk membantu mempermudah melakukan klasifikasi, klasterisasi dan prediksi. Kelebihan pembelajaran mesin adalah dapat belajar dari contoh bukan dengan aturan. Untuk sebuah tugas yang diberikan, contoh diberikan dalam bentuk masukan (fitur) dan keluaran (label). Komputer kemudian menentukan cara melakukan pemetaan dari fitur ke label untuk membuat model yang akan menghasilkan informasi (Rajkomar et al., 2019). Implementasi pembelajaran mesin dalam bidang kesehatan hingga saat ini sudah banyak diterapkan di dunie medis seperti perekaman data kesehatan, pencitraan medis hingga bidang genetic seperti pengenalan DNA(Habehh & Gohel, 2021). Namun ada beberapa tantangan dalam mengimplementasikan pembelajaran mesin di bidang medis seperti kualitas data dan aplikasi pembelajaran mesin yang mudah digunakan oleh pengguna (Alanazi, 2022).

Ada banyak algoritma yang digunakan untuk pemodelan pembelajaran mesin di bidang medis seperti naïve bayes, SVM, Decision Tree, KNN dan Random Forest. Dari beberapa hasil penelitian, random forest terbukti lebih akurat dibanding algoritma lain seperti Naïve Bayes, KNN, Decision Tree dan SVM (Gaurav et al., 2023). Tujuan dari penelitian ini adalah untuk membantu tenaga medis seperti dokter untuk melakukan klasifikasi apakah pasien memiliki penyakit jantung atau tidak dengan model pembelajaran mesin algoritma random forest.

TINJAUAN PUSTAKA

Penelitian Terkait

Penelitian yang terkait dengan penyakit jantung sudah dilakukan menggunakan algoritma K-Nearest Neighbor (KNN), dimana tingkat akurasi mencapai 86.8% (Garg et al., 2021). Prediksi penyakit jantung juga dilakukan dengan menggunakan fitur yang berbeda seperti usia, nyeri dada, tekanan darah, kolesterol dan fitur lainnya(Pal & Parija, 2021). Selain memprediksi penyakit jantung, machine lerning juga dapat memprediksi penyakit diabetes dengan menggunakan beberapa algoritma yaitu Support Vector Machine, Decision Tree, Logistic Regression, Random Forest,

Jurnal Minfo Polgan
Volume 12, Nomor 2, November 2023 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13214 p-ISSN : 2089-9424

Neural Network dan Naïve Bayes (Xie et al., 2019). Untuk memprediksi resiko diabetes tipe-2 di jepang, peneliti menggunakan model random forest (Vlachas et al., 2022).

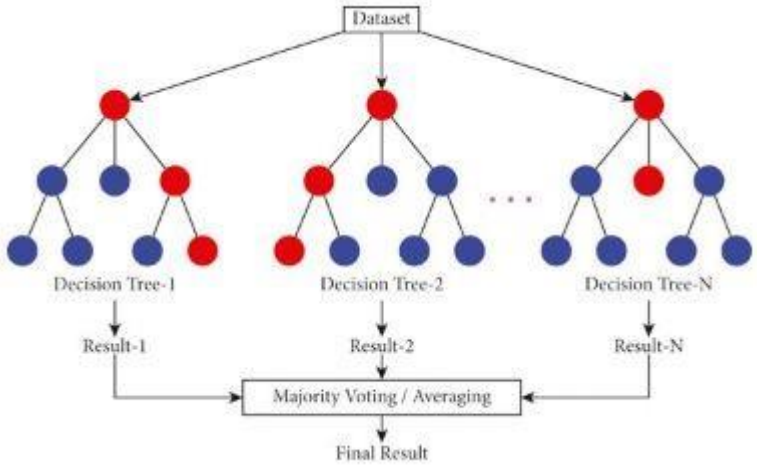
Dalam lingkup kepentingan medis dan lebih luas, machine learning dapat digunakan untuk memprediksi berbagai penyakit seperti penyakit liver, diabetes dan jantung dengan menggabung kan model random forest dan transfer learning (Vlachas et al., 2022).

Penyakit Jantung

Penyakit jantung adalah suatu kondisi dimana jantung tidak dapat melaksanakan fungsinya dengan baik, sehingga fungsi jantung sebagai pemompa darah dan oksigen ke seluruh tubuh terganggu. Terganggunya proses peredaran oksigen dan darah tersebut dapat disebabkan karena otot jantung yang melemah, adanya celah antara serambi kiri dan serambi kanan yang mengakibatkan darah bersih dan darah kotor tercampur (Anies & Andin, 2015). Penyakit jantung merupakan penyakit tidak menular terbanyak di dunia, WHO merilis bahwa pada tahun 2017 menyebabkan kematian hinga 17.8 juta jiwa di seluruh dunia (WHO) (World Health Organization, 2015).

Random Forest

Random forest diperkenalkan pertama kali oleh Leo Breiman di dalam jurnalnya yang berjudul Random Forests tahun 2001. Random Forest adalah model yang dapat membentuk sejumlah pohon (tree) dimana setiap pohon yang dihasilkan berasal dari sejumlah data training. Setiap pohon yang dibentuk menghasilkan prediksi kelas dan prediksi kelas dengan vote terbanyak (majority vote) yang akan menjadi acuan untuk model prediksi (Breiman, 2001) seperti gambar di bawah ini :



Setiap pohon dibentuk menggunakan algoritma decision tree dengan menghitung gain dan entropy.

$$\text{Entropy} = \sum_{i=1}^{k} -p_i \log_2 p_i$$

METODE PENELITIAN

CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP-DM (Cross-Industry Standard Process for Data Mining) digunakan sebagai kerangka kerja dalam menerapkan proses data mining (Daniel T. Larose, n.d.) adapun prosesnya yaitu :

1. Business Understanding Phase

Tahap ini dimulai dengan memahami tujuan bisnis yang ingin dicapai. Identifikasi permasalahan yang ingin dipecahkan dan tujuan proyek dari perspektif bisnis. Ini melibatkan komunikasi dengan para pemangku kepentingan untuk memahami kebutuhan mereka..

2. Data Understanding Phase

Jurnal Minfo Polgan
Volume 12, Nomor 2, November 2023 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13214 p-ISSN : 2089-9424

Pada tahap ini, data yang tersedia dikumpulkan. Proses eksplorasi data dilakukan untuk memahami karakteristik data, seperti sifat, struktur, dan kualitasnya. Tujuan dari tahap ini adalah mendapatkan pemahaman yang lebih dalam tentang data yang akan digunakan dalam proyek.

3. Data Preparation Phase

Data yang telah dikumpulkan diolah pada tahap ini. Langkah-langkah ini meliputi pembersihan data, penanganan nilai yang hilang, transformasi data, serta pengolahan data agar siap untuk proses pemodelan. Hal ini melibatkan teknik-teknik seperti pembersihan data, normalisasi, transformasi variabel, dan teknik lainnya untuk mempersiapkan data.

4. Modelling Phase

Setelah persiapan data, model-model untuk analisis atau prediksi dibangun. Proses pemodelan melibatkan pemilihan teknik-teknik analisis yang sesuai dan penggunaan algoritma atau model yang cocok dengan data yang dimiliki.

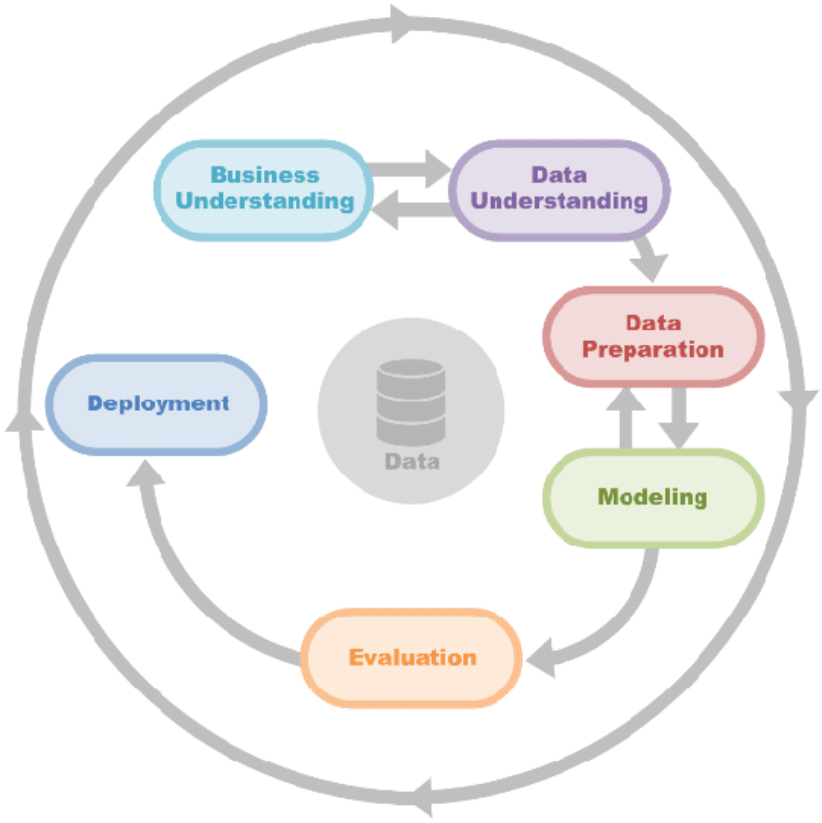
5. Evaluation Phase

Model-model yang telah dibuat dievaluasi untuk memastikan bahwa mereka sesuai dengan tujuan bisnis yang telah ditetapkan. Proses evaluasi melibatkan pengujian kinerja model, membandingkan hasil prediksi dengan data aktual, serta memeriksa apakah model memenuhi standar yang ditetapkan.

6. Deployment Phase

Setelah model dievaluasi dan disetujui, langkah terakhir adalah mengimplementasikan solusi yang ditemukan. Ini melibatkan penyampaian hasil analisis ke dalam pengambilan keputusan bisnis atau sistem yang ada.

Setelah tahap implementasi, ada kemungkinan untuk kembali ke tahap awal atau tahap mana pun, terutama jika ada perubahan dalam kebutuhan bisnis, data baru yang tersedia, atau perubahan lain yang mempengaruhi proyek. CRISP-DM merupakan proses yang iteratif, memungkinkan untuk kembali ke tahap sebelumnya jika diperlukan untuk memastikan bahwa hasil akhir sesuai dengan tujuan bisnis yang diinginkan.

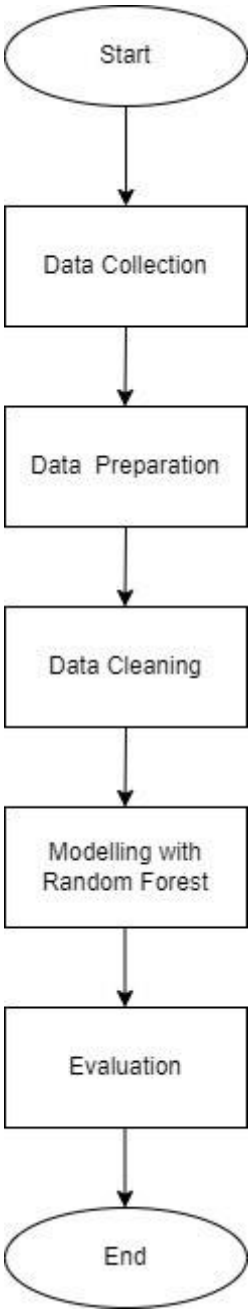


Gambar 1 CRISP-DM Process

Jurnal Minfo Polgan
Volume 12, Nomor 2, November 2023 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13214 p-ISSN : 2089-9424

Flowchart Pemodelan Random Forest

Proses dimulai dari penyiapan data-data yang dibutuhkan untuk pemodelan. Dari data yang sudah ada, tentunya data kemungkinant tidak bersih sehingga harus dilakukan pembersihan data. Setelah data sudah bersih, maka dilakukan pemilihan fitur dan klasifikasi data. Dikarenakan algoritma yang digunakan adalah random forest, maka setiap fitur harus dilakukan klasifikasi untuk kemudian akan diproses ke dalam model random forest. Data yang diproses otomatis akan terbentuk menjadi dua, yaitu data training dan data testing. Setelah hasil pemodelan, maka dilakuan evaluasi terhadap hasil seperti akurasi dan presisi. Detail proses digambarkan dalam flowchart di bawah ini :



Gambar 2 Proses Pemodelan

Jurnal Minfo Polgan
Volume 12, Nomor 2, November 2023 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13214 p-ISSN : 2089-9424

HASIL DAN PEMBAHASAN

Dibutuhkan satu pemodelan machine learning yang akan digunakan untuk memprediksi apakah pasien memiliki kemungkinan penyakit jantung atau tidak. Machine learning akan membantu dokter dalam mempercepat dan mempermudah mendapatkan diagnosa sementara pasien. Hasil diagnosa bersifat untuk mendukung keputusan karena tentunya perlu ada pemeriksaan lebih lanjut secara medis oleh dokter spesialis jantung.

Data yang diambil merupakan data indicator penyakit jantung yang bersumber dari Behavioral Risk Factor Surveillance System (BRFSS). Organisasi pemerintah ini melakukan survey kesehatan via telepon setiap tahun di amerika yang berkaitan dengan resiko kesehatan, kondisi kesehatan kronis dan digunakan untuk layanan pencegahan. Data diambil sampai dengan tahun 2015 berjumlah 253.580 baris data dengan 22 fitur.

Dari data yang sudah ada keseluruhan yaitu total 22 fitur, adapun masing-masing fitur itu adalah :

Table 1 Deskripsi Fitur

Fitur	Deskripsi
HighBP	Memiliki tekanan darah tinggi
HighCholesterol	Memiliki kolesterol tinggi
CholCheck	Melakukan pengecekkan rutin kolesterol
BMI	Body Mass Index
Smoker	Perokok
Stroke	Memiliki riwayat stroke
Diabetes	Memiliki diabetes tipe 2
PhysActivity	Aktifitas fisik yang dilakukan sebulan terakhir
Fruits	Mengkonsumsi buah-buahan minimal 1 buah perhari
Veggies	Mengkonsumsi sayuran minimal 1 buah perhari
HvyAlcoholConsump	Peminum alcohol berat
AnyHealthcare	Memiliki akses kesehatan atau asuransi
NoDocbcCost	Tidak memiliki biaya perawatan kesehatan dalam setahun terakhir
GenHlth	Tingkatan kesehatan
MentHlth	Kondisi kesehatan mental dalam 30 hari terakhir
PhysHlth	Kondisi kesehatan fisik dalam 30 hari terakhir
DiffWalk	Kondisi kesulitan berjalan atau menaiki tangga
Sex	Jenis kelamin
Age	Usia
Education	Tingkat pendidikan
Income	Tingkat pendapatan
HeartDiseaseorAttack	Mengalami penyakit jantung atau tidak

Dari 22 fitur, 1 kelas adalah fitur HeartDiseaseorAttack dimana fitur sudah diklasifikasikan menjadi 0 dan 1, 0 adalah tidak dan 1 adalah ya. Untuk penjelasan dari klasifikasi fitur adalah dijelaskan tabel di bawah ini :

Jurnal Minfo Polgan
Volume 12, Nomor 2, November 2023
DOI : https://doi.org/10.33395/jmp.v12i2.13214
e-ISSN : 2797-3298
p-ISSN : 2089-9424

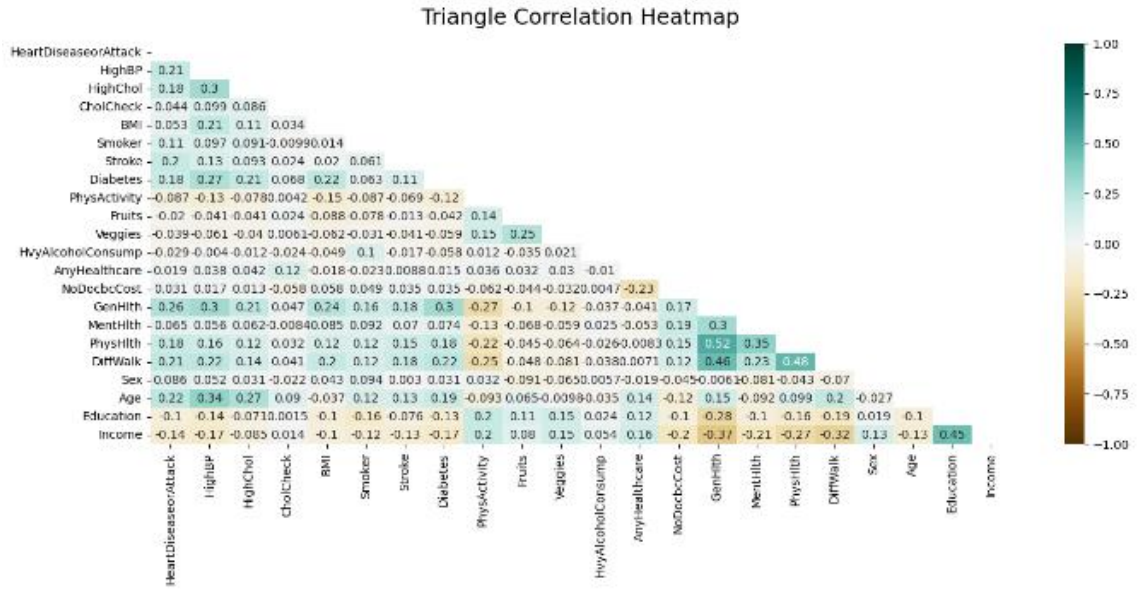
Table 2 Klasifikasi Fitur

Fitur	Klasifikasi
HighBP	0=tidak, 1=ya
HighCholesterol	0=tidak, 1=ya
CholCheck	0=tidak, 1=ya
BMI	0-20, 21-30, 31-40, 41-50, 51-60, 61-80, >80
Smoker	0=tidak, 1=ya
Stroke	0=tidak, 1=ya
Diabetes	0=tidak, 1=ya
PhysActivity	0=tidak, 1=ya
Fruits	0=tidak, 1=ya
Veggies	0=tidak, 1=ya
HvyAlcoholConsump	0=tidak, 1=ya
AnyHealthcare	0=tidak, 1=ya
NoDocbcCost	0=tidak, 1=ya
GenHlth	1=excellent, 2=very good, 3=good, 4=fair, 5=poor
MentHlth	0 hari 1-10 hari 11-20 hari 21-30 hari
PhysHlth	0 hari 1-10 hari 11-20 hari 21-30 hari
DiffWalk	0=tidak, 1=ya
Sex	0=perempuan, 1=laki-laki
Age	18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, >79
Education	1=tidak pernah sekolah, 2= sekolah dasar 3=sekolah menengah pertama 4=sekolah menengah atas 5=diploma 6=sarjana
Income	1: < $ 10000 2: $ 10000 - $ 15000 3: $ 15000 - $ 20000 4: $ 20000 - $ 25000 5: $ 25000 - $ 35000 6: $ 35000 - $ 50000 7: $ 50000 - $ 75000 8: > $ 75000
HeartDiseaseorAttack	0=tidak, 1=ya

Jurnal Minfo Polgan
Volume 12, Nomor 2, November 2023 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13214 p-ISSN : 2089-9424

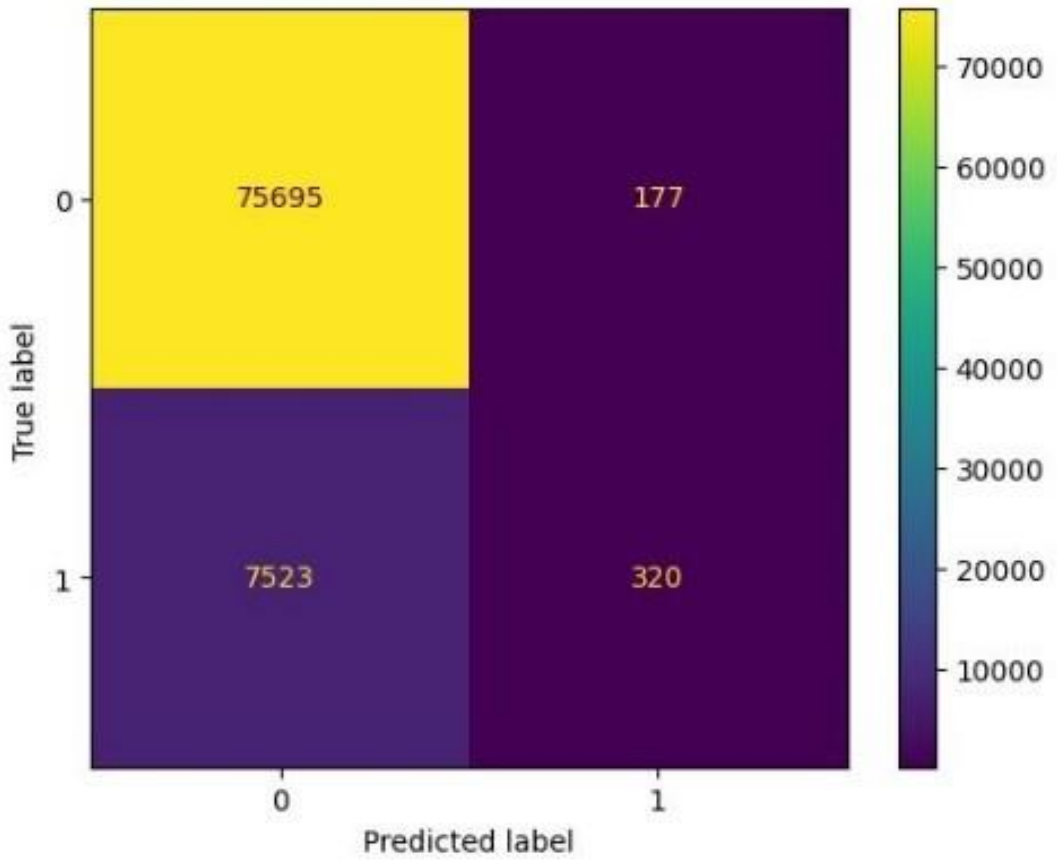
Dari data di atas kemudian dikonversi ke dalam triangle correlaction heatmap yang berfungsu untuk memetakan seberapa kuat korelasi antar fitur seperti di bawah ini:

Table 3 Triangle Correlation Heatmap



Pengujian model menggunakan Confusion Matrix menampilkan hasil seperti di bawah ini :

Table 4 Confusion Matrix



Jurnal Minfo Polgan
Volume 12, Nomor 2, November 2023 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13214 p-ISSN : 2089-9424

Akurasi, recall dan presisi menggunakan Python ditunjukkan oleh tabel dibawah ini:

Table 5 Presisi, Akurasi dan Recall

	precision	recall	f1-score	support
0	0.91	1.00	0.95	75872
1	0.64	0.04	0.08	7843
accuracy			0.91	83715
macro avg	0.78	0.52	0.51	83715
weighted avg	0.88	0.91	0.87	83715

KESIMPULAN

Dari hasil evaluasi, random forest terbukti dapat memprediksi penyakit jantung dengan akurasi dan presisi 91%. Dengan melihat hasil pengujian, dapat disimpulkan bahwa algoritma random forest memiliki akurasi yang cukup tinggi dan mampu memprediksi penyakit jantung berdasarkan banyak paramater. Dengan pemodelan ini, maka akan membantu tenaga medis untuk lebih mudah sebagai alat bantu deteksi lebih dini.

REFERENSI

Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. In *Informatics in Medicine Unlocked* (Vol. 30). Elsevier Ltd. https://doi.org/10.1016/j.imu.2022.100924

Anies, & Andin. (2015). *Kolesterol dan Penyakit Jantung Koroner*.

Breiman, L. (2001). *Random Forests* (Vol. 45).

Daniel T. Larose. (n.d.). *Data Mining Methode and Models*.

Garg, A., Sharma, B., & Khan, R. (2021). Heart disease prediction using machine learning techniques. *IOP Conference Series: Materials Science and Engineering*, *1022*(1). https://doi.org/10.1088/1757-899X/1022/1/012046

Gaurav, K., Kumar, A., Singh, P., Kumari, A., Kasar, M., & Suryawanshi, T. (2023). Human Disease Prediction using Machine Learning Techniques and Real-life Parameters. *International Journal of Engineering, Transactions B: Applications*, *36*(6), 1092–1098. https://doi.org/10.5829/ije.2023.36.06c.07

Habehh, H., & Gohel, S. (2021). Machine Learning in Healthcare. *Current Genomics*, *22*(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

Pal, M., & Parija, S. (2021). Prediction of Heart Diseases using Random Forest. *Journal of Physics: Conference Series*, *1817*(1). https://doi.org/10.1088/1742-6596/1817/1/012009

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/nejmra1814259

Vlachas, C., Damianos, L., Gousetis, N., Mouratidis, I., Kelepouris, D., Kollias, K.-F., Asimopoulos, N., & Fragulis, G. F. (2022). Random forest classification algorithm for medical industry data. *SHS Web of Conferences*, *139*, 03008. https://doi.org/10.1051/shsconf/202213903008

World Health Organization. (2001). *Cardiovascular diseases (CVDs)*. Https://Www.Who.Int/News-Room/Fact-Sheets/Detail/Cardiovascular-Diseases-(Cvds).

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, *16*(9). https://doi.org/10.5888/pcd16.190109