

PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN METODE RANDOM FOREST DAN PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA)

Nur Halizah Alfajr^{1*}, Sofi Defiyanti²

^{1,2}Universitas Singaperbangsa Karawang, Jl.HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat 41361, Telp. (0267) 641177

Received: 24 Juli 2024

Accepted: 5 Oktober 2024

Published: 12 Oktober 2024

Keywords:

Klasifikasi, Prediksi,
Penyakit Jantung, Random
Forest Classifier

Correspondent Email:

nurhlzh2403@gmail.com

Abstrak. Penyakit jantung merupakan masalah kesehatan masyarakat yang signifikan dan menjadi penyebab kematian utama di seluruh dunia. Faktor risiko seperti hipertensi, diabetes, obesitas, gaya hidup tidak aktif, merokok, dan faktor genetik berkontribusi terhadap perkembangan penyakit ini. Penelitian ini bertujuan untuk mengembangkan model prediksi penyakit jantung menggunakan metode Random Forest. Dataset yang digunakan berasal dari UCI Machine Learning Repository dengan sampel 1026 pasien yang memiliki berbagai indikator kesehatan. Proses yang dilakukan meliputi tahapan dalam *knowledge discovery in database (KDD)*, yaitu *data selection*, *preprocessing*, *transformation*, *data mining* dan *evaluation*. Hasil penelitian menunjukkan bahwa model dengan 100 pohon keputusan mencapai akurasi 0.9823. Evaluasi menggunakan *confusion matrix* dan *classification report* menunjukkan metode Random Forest memberikan akurasi 98%, presisi 100%, recall 96%, dan F1-score 98%. Kesimpulannya, metode Random Forest efektif untuk memprediksi penyakit jantung, dengan fitur *thal* yang sangat mempengaruhi akurasi model.

Abstract. Heart disease is a significant public health issue and the leading cause of death worldwide. Risk factors such as hypertension, diabetes, obesity, sedentary lifestyle, smoking, and genetic factors contribute to the development of heart disease. This study aims to develop a heart disease prediction model using the Random Forest method. The dataset used comes from the UCI Machine Learning Repository, containing data from 1026 patients with various health features. The methods used include the stages of knowledge discovery in databases (KDD), namely data selection, preprocessing, transformation, data mining, and evaluation. The study results show that the model with 100 decision trees achieved an accuracy of 0.9823. Further evaluation using the confusion matrix and classification report indicates that the Random Forest method provides 98% accuracy, 100% precision, 96% recall, and a 98% F1-score. In conclusion, the Random Forest method is effective in predicting heart disease, with features such as *thal* having a significant impact on model accuracy.

1. PENDAHULUAN

Penyakit jantung adalah salah satu masalah kesehatan utama yang menyebabkan kematian terbanyak di dunia. Faktor-faktor risiko seperti tekanan darah tinggi, diabetes, obesitas, gaya hidup tidak sehat, merokok, dan faktor genetik

dapat memicu penyakit jantung [1]. Deteksi dini dan pengendalian faktor risiko sangat penting untuk mencegah dan mengelola penyakit ini [2].

Algoritma pembelajaran mesin, terutama metode Random Forest, semakin populer dalam

mendeteksi dan memprediksi penyakit jantung [3]. Pembelajaran mesin, bagian dari kecerdasan buatan, memungkinkan komputer untuk belajar dan membuat keputusan tanpa harus diprogram ulang secara eksplisit [4]. ini memberi mesin kemampuan untuk berperilaku mengambil keputusan dan menyesuaikan diri dengan perubahan. Machine learning memberikan hasil yang tepat dengan teknik yang cerdas untuk menangani data besar [5].

Beberapa penelitian telah dilakukan mengenai prediksi penyakit jantung menggunakan berbagai metode yang telah dilakukan sebelumnya. Studi yang dilakukan oleh [6] membandingkan model Decision Tree, Naïve Bayes, dan Random Forest untuk prediksi klasifikasi penyakit jantung, menemukan nilai akurasi metode Decision Tree 0.71%, Naïve Bayes 0.72%, dan Random Forest 0.75%.

Studi yang dilakukan oleh [7] mencoba memprediksi penyakit gagal jantung dengan menggunakan Random Forest, dan menemukan nilai akurasi 82.6087%.

Studi [8] membandingkan kinerja algoritma untuk memprediksi penyakit jantung dengan metode data mining. Metode Random Forest menghasilkan nilai prediksi yang akurat dengan presentase sebesar 83%.

Penelitian – penelitian ini membandingkan kinerja model Random Forest dengan studi sebelumnya tentang prediksi penyakit jantung. Penelitian ini juga menggunakan Principal Component Analysis (PCA) untuk meningkatkan kinerja model dan menunjukkan keunggulan Random Forest dalam akurasi prediksi.

2. TINJAUAN PUSTAKA

Data mining adalah proses analitis yang bertujuan mengekstraksi informasi berharga dari sejumlah besar data. Klasifikasi adalah salah satu metode dalam data mining yang berguna untuk membuat prediksi menggunakan data yang padat. Dalam penelitian ini, algoritma Random Forest Classifier digunakan [9].

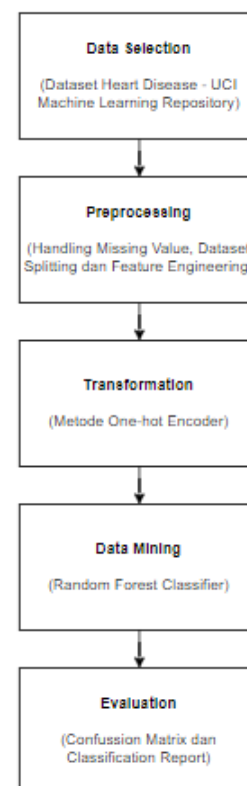
Penyakit jantung merupakan penyebab kematian utama, termasuk penyakit jantung koroner, serangan jantung, dan gagal jantung. Identifikasi dan prediksi risiko penyakit jantung penting untuk pengobatan yang tepat dan pencegahan komplikasi. Algoritma pembelajaran mesin mampu menganalisis data

medis untuk menemukan pola yang berhubungan dengan risiko penyakit jantung dan memberikan rekomendasi klinis yang lebih baik [10].

Random forest adalah salah satu algoritma pembelajaran mesin yang efektif, menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi. Algoritma ini menggunakan sebagian kecil dari atribut dan subset data secara acak untuk melatih setiap pohon, kemudian menggabungkan hasil prediksi untuk menentukan klasifikasi akhir [11]. Metode ensemble ini mengurangi risiko overfitting dan meningkatkan kemampuan generalisasi model, membuat Random Forest menjadi pilihan yang tepat untuk tugas klasifikasi.

3. METODE PENELITIAN

Penelitian ini menggunakan metodologi *Knowledge Discovery in Database (KDD)*, yang meliputi beberapa tahapan, yaitu *data selection*, *preprocessing*, *transformation*, *data mining* dan *evaluation*, sebagaimana ditunjukkan pada gambar 1.



Gambar 1. Tahapan Penelitian

3.1. Data Selection

Data yang digunakan oleh penulis terdiri dari 1026 record dan terdiri dari 14 fitur. Tabel 1 menggambarkan atribut atau indikator kumpulan data yang akan digunakan.

Tabel 1. Atribut Dataset

Atribut	Keterangan
age	Usia pasien dalam tahun
sex	Jenis kelamin pasien (1 = laki-laki, 0 = perempuan).
cp	Tipe nyeri dada (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic).
trestbps	Tekanan darah istirahat dalam mm Hg pada saat masuk rumah sakit.
chol	Kadar kolesterol serum dalam mg/dl.
fbs	Gula darah puasa > 120 mg/dl (1 = benar, 0 = salah).
restecg	Hasil elektrokardiografi istirahat (0 = normal, 1 = memiliki kelainan gelombang ST-T, 2 = hipertrofi ventrikel kiri).
thalach	Detak jantung maksimum yang dicapai.
exang	Angina yang diinduksi oleh olahraga (1 = ya, 0 = tidak).
oldpeak	Depresi ST yang diinduksi oleh olahraga relatif terhadap istirahat.
slope	Kemiringan segmen ST pada puncak latihan (0 = upsloping, 1 = flat, 2 = downsloping).
ca	Jumlah pembuluh darah utama (0-3) yang diwarnai oleh fluoroskopi.
thal	Kondisi thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect).
target	Diagnosis penyakit jantung (1 = memiliki penyakit jantung, 0 = tidak memiliki penyakit jantung).

Tabel 1 merupakan atribut data set yang terdiri 14 atribut data yang meliputi age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target.

3.2. Preprocessing

Selanjutnya adalah tahap *preprocessing*, pada tahap ini data mentah akan dilakukan

proses *handling missing value* untuk memastikan apakah ada data yang kosong atau data yang salah dimasukkan dengan menggunakan library pandas. Setelah itu data akan melalui tahap *dataset splitting* dan *feature engineering*.

3.3. Transformation

Transformasi data adalah proses mengubah data dari format aslinya ke format yang lebih cocok, misalnya mengubah data nominal menjadi data numerik. Tahapan ini bertujuan untuk menyesuaikan data terpilih sehingga memudahkan proses penambangan data [12]. Pada tahap ini, akan dilakukan encode variabel kategorikal dengan bentuk yang dapat dimengerti oleh algoritma Random Forest Classifier.

3.4. Data Mining

Langkah pertama dalam tahap klasifikasi dengan metode random forest adalah membuka data yang diekstraksi di Google Collab menggunakan library pandas. Untuk mendapatkan parameter yang optimal dalam hal ini, penulis menggunakan teknik menentukan jumlah decision tree yang akan digunakan dalam ensemble dan menetapkan nilai seed untuk memastikan hasil yang konsisten.

3.5. Evaluation

Langkah terakhir dalam menilai kinerja algoritma klasifikasi yang digunakan dalam penelitian ini adalah evaluasi. Ada empat metrik yang digunakan untuk mengukur kinerja model, yaitu [13] :

1. *Accuracy* yaitu proporsi data yang diprediksi dengan benar dari keseluruhan data.
2. *Recall*, metrik yang menunjukkan seberapa banyak data dari kelas positif yang diprediksi dengan benar
3. *Precision*, mengukur seberapa akurat model dalam memprediksi data dari kelas positif.
4. *F-measure*, nilai yang dihasilkan dari rata-rata gabungan antara precision dan recall.
5. *Confusion Matrix*, alat untuk mengevaluasi kinerja model klasifikasi dalam machine learning. Matriks ini membandingkan hasil prediksi model dengan nilai aktual dalam bentuk matriks, yang terdiri dari empat komponen utama:

True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN)

4. HASIL DAN PEMBAHASAN

Hasil dan pembahasan prediksi penyakit jantung menggunakan metode Random Forest dan penerapan PCA sebagai berikut.

4.1. Preprocessing

4.1.1. Handling Missing Value

Tahap melakukan *handling missing value* menggunakan metode `isnull` dan `sum` yaitu mengembalikan dataframe yang berisi nilai boolean (True atau False) dan menjumlahkan nilai boolean dalam dataframe. Karena nilai True dianggap 1 dan nilai False dianggap 0, hasilnya adalah jumlah nilai yang hilang untuk setiap variabel dalam dataframe.

```
age      0
sex      0
cp       0
trestbps 0
chol     0
fbs      0
restecg  0
thalach  0
exang    0
oldpeak  0
slope    0
ca       0
thal     0
target   0
dtype: int64
```

Gambar 2. Deskripsi Missing Value

Hasil dari tahapan ini adalah pada Gambar 2 terlihat bahwa data telah terisi secara keseluruhan (tidak ditemukan missing value).

4.1.2. Dataset Splitting

Setelah tahap *handling missing value*, tahap berikutnya adalah melakukan pembagian data serta memisahkan antara data train dan data test.

```
((686, 13), (339, 13))
```

Gambar 3. Pembagian Data

Hasil tahapan ini menunjukkan data train (latih) sebanyak 67% dan data test (uji) sebanyak 33%.

4.1.3. Feature Engineering

Pada tahap ini Fungsi `X_train.dtypes` digunakan untuk menampilkan tipe data dari setiap fitur (kolom) dalam DataFrame `X_train` dimana tahap ini bertujuan untuk mengubah data mentah menjadi fitur berguna serta memahami model dengan lebih baik.

```
age      int64
sex      int64
cp       int64
trestbps int64
chol     int64
fbs      int64
restecg  int64
thalach  int64
exang    int64
oldpeak  float64
slope    int64
ca       int64
thal     int64
dtype: object
```

Gambar 4. Hasil Featurig Engineering

Hasil dari tahap ini menunjukkan dataset `X_train` terdiri dari atribut-atribut numerik dengan mayoritas berupa bilangan bulat (integer) dan satu atribut (`oldpeak`) berupa bilangan decimal (float).

4.2. Transformation

4.2.1. Encode Variabel Kategorikal

Pada tahap ini, penulis menggunakan `one-hot encoding` yang berfungsi untuk mengubah atribut kategori menjadi bentuk numerik. Kemudian dilakukan transformasi data uji menggunakan `pandas.get_dummies`, memastikan semua kolom kategorikal dalam data uji sudah di encode, memastikan urutan kolom yang sama dengan data latih serta menggabungkan dengan fitur yang lainnya.

Untuk data training

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
866	41	1	2	112	250	0	1	179	0	0.0	2	0	2
749	58	1	1	125	220	0	1	144	0	0.4	1	4	3
212	50	1	0	150	243	0	0	128	0	2.6	1	0	3
79	52	1	1	134	201	0	1	158	0	0.8	2	1	2
148	52	1	3	152	298	1	1	178	0	1.2	1	0	3

Gambar 5. Dataset Training

Untuk data testing

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
527	62	0	0	124	209	0	1	163	0	0.0	2	0	2
359	53	0	2	128	216	0	0	115	0	0.0	2	0	0
447	55	1	0	160	289	0	0	145	1	0.8	1	1	3
31	50	0	1	120	244	0	1	162	0	1.1	2	0	2
621	48	1	0	130	256	1	0	150	1	0.0	2	2	3

Gambar 6. Dataset Testing

Hasil dari tahap ini menunjukkan data pada atribut yang mengandung kategorikal berubah menjadi numerik.

4.3. Data Mining

Penulis kemudian memulai model Random Forest yang akan digunakan untuk membuat prediksi berdasarkan karakteristik data yang telah diubah.

Tabel 2. Model 10 Decision Tree

Random Forest Classifier model with parameters n_estimators=10
<pre>from sklearn.ensemble import RandomForestClassifier rfc = RandomForestClassifier(n_estimators=10, random_state=0) rfc.fit(X_train, y_train) y_pred = rfc.predict(X_test)</pre>
<pre>from sklearn.metrics import precision_score print('Model accuracy score with 10 decision- trees : {0:0.4f}'.format(precision_score(y_test, y_pred)))</pre>
Model accuracy score with 10 decision-trees : 1.0000

Hasil model nilai akurasi klasifikasi Random Forest dengan parameter n_estimator = 10 adalah 1.000.

Tabel 3. Model 100 Decision Tree

Random Forest Classifier model with parameters n_estimators=100
<pre>rfc_1000 = RandomForestClassifier(n_estimators=1000, random_state=0) rfc_1000.fit(X_train, y_train) y_pred_1000 = rfc_1000.predict(X_test)</pre>
<pre>print('Model accuracy score with 100 decision- trees : {0:0.4f}'.format(accuracy_score(y_test, y_pred_1000)))</pre>
Model accuracy score with 100 decision-trees : 0.9823

Hasil model nilai akurasi klasifikasi Random Forest dengan parameter n_estimator = 100 adalah 0.9823.

Selanjutnya, penulis akan menggunakan Random Forest Classifier dengan n_estimators = 100, yang berarti akan menggunakan 100 pohon keputusan dalam pembentukan model. Tetapi sebelum itu, penulis akan menampilkan daftar skor penting fitur yang dapat digunakan untuk menganalisis dan memahami kontribusi masing – masing fitur terhadap model.

thal	0.133865
cp	0.128325
oldpeak	0.123375
ca	0.115472
thalach	0.107888
age	0.086074
chol	0.083262
trestbps	0.070425
slope	0.046238
exang	0.042454
sex	0.035124
restecg	0.017505
fbs	0.009993
dtype:	float64

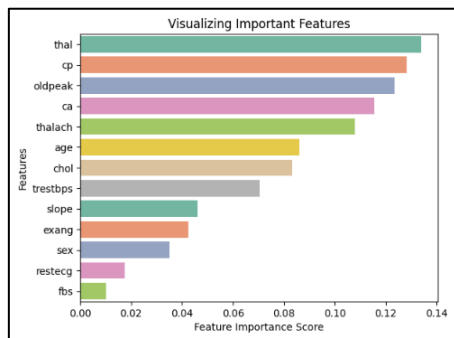
Gambar 7. Hasil Skor Fitur

Tabel 4. Model Akurasi Seluruh Fitur

Akurasi dengan semua fitur
<pre>from sklearn.tree import DecisionTreeClassifier from sklearn.metrics import classification_report, confusion_matrix</pre>
<pre>dt= DecisionTreeClassifier(random_state=0) dt.fit(X_train, y_train) y_pred = dt.predict(X_test)</pre>
<pre>from sklearn.metrics import accuracy_score print('Model accuracy score with All Features : {0:0.4f}'.format(accuracy_score(y_test, y_pred)))</pre>
Model accuracy score with All Features : 0.9617

Hasil model nilai akurasi dengan memperhitungkan semua fitur adalah 0.9617.

Gambar 8 menunjukkan hasil visualisasi fitur penting.



Gambar 8. Visualisasi Fitur Penting

Hasil menunjukkan fitur fbs tergolong kedalam fitur yang tidak penting karena berada diposisi terakhir. Selanjutnya, fitur fbs akan dihapus dengan menerapkan Principal Component Analysis (PCA) untuk membuktikan apakah klasifikasi Random Forest lebih akurat.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Gambar 9. Dataset tanpa fitur Fbs

Tabel 5. Model Akurasi Fbs Dihapus

Akurasi dengan menghapus fitur fbs

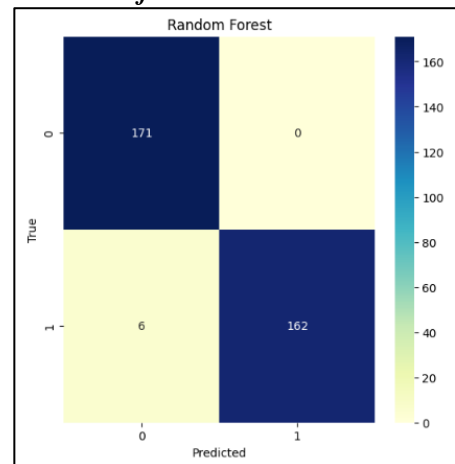
```
clf = RandomForestClassifier(n_estimators=100,
random_state=0)
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print('Model accuracy score with fbs variable
removed : {0:0.4f}'.
format(accuracy_score(y_test, y_pred)))
```

Model accuracy score with fbs variable removed : 0.9823

Hasil model nilai akurasi dengan penghapusan variabel fbs yaitu sebesar 0.9823.

4.4. Evaluation

4.4.1. Confusion Matrix



Gambar 10. Visualisasi Confusion Matrix

Kelas 0 (Kelas Aktual 0)

- Diprediksi sebagai Kelas 0: 171 kali (*True Positive*)
- Diprediksi sebagai Kelas 1: 0 kali (*False Negative*)

Kelas 1 (Kelas Aktual 1)

- Diprediksi sebagai Kelas 0: 6 kali (*False Positive*)
- Diprediksi sebagai Kelas 1: 162 kali (*True Positive*)

4.4.2. Classification Report

	precision	recall	f1-score	support
0	0.97	1.00	0.98	171
1	1.00	0.96	0.98	168
accuracy			0.98	339
macro avg	0.98	0.98	0.98	339
weighted avg	0.98	0.98	0.98	339

Gambar 11. Visualisasi Classification Report

Dari Gambar 11 yang tertera di atas, terlihat bahwa metode Random Forest memberikan hasil yaitu sebesar akurasi 98%, presisi 100%, recall 96%, dan F1-score 98%.

Dalam penelitian ini, metode Random Forest diterapkan untuk memprediksi penyakit jantung, dengan menggunakan Principal Component Analysis (PCA) untuk mengoptimalkan fitur. Proses penelitian yang terdiri dari tahap *data selection*, *preprocessing*, *transformation*, *data mining* dan *evaluation* di mana data dibagi menjadi dua bagian: data train

(67%) dan data test (33%), lalu dilakukan proses transformation dan selanjutnya diproses menggunakan algoritma Random Forest. Hasil confusion matrix menunjukkan bahwa model mampu memprediksi kelas 0 sebanyak 171 kali dan kelas 1 sebanyak 162 kali dengan tingkat kesalahan yang sangat rendah, yaitu hanya 6 kali False Positive. Classification Report menunjukkan bahwa model menghasilkan akurasi 98%, presisi 100%, recall 96%, dan F1-score 98%.

Secara keseluruhan, penelitian ini berhasil menunjukkan bahwa metode Random Forest dengan penerapan PCA menghasilkan model prediksi penyakit jantung yang sangat akurat. Hasil evaluasi menunjukkan bahwa model memiliki tingkat akurasi, presisi, recall, dan F1-score yang sangat tinggi, menjadikan metode ini sebagai pilihan yang baik untuk prediksi penyakit jantung.

5. KESIMPULAN

Penelitian telah mencapai beberapa kesimpulan berikut:

1. Tahap prediksi penyakit jantung menggunakan algoritma Random Forest Classifier terdiri dari beberapa tahap: *Data Selection*, *Preprocessing*, *Data Transformation*, *Data Mining* dan *Evaluation*.
2. Hasil prediksi menyatakan bahwa akurasi sebesar 98%, presisi 100%, recall 96% dan F1-score 98%, data testing yang digunakan sebanyak 33% dari 1026 data.
3. Model Random Forest membantu menemukan fitur – fitur penting saja, membangun model menggunakan fitur-fitur ini dan melihat pengaruhnya terhadap akurasi. Fitur yang paling berpengaruh adalah *thal* dan fitur yang paling tidak berpengaruh adalah *fbs*.
4. Keakuratan model dengan variabel *fbs* dihapus adalah 0.9823. Keakuratan model dengan memperhitungkan semua variabel dengan $n_estimator = 100$ adalah 0.9617. Jadi, kita dapat melihat bahwa akurasi model dari hilangnya variabel *fbs* terjadi perubahan yaitu mengalami kenaikan nilai akurasi atau bisa dibilang fitur *fbs* mempengaruhi keakuratan model setelah dihilangkan. Hal ini bisa saja dilakukan juga pada fitur lain, tetapi dapat mengganggu isi dari dataset tersebut.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak terkait yang berpartisipasi dalam penelitian ini.

DAFTAR PUSTAKA

- [1] “Cardiovascular disease (CVDs),” World Health Organization, 11 Juni 2021. [Online]. [Accessed 11 Juli 2024].
- [2] A. M. Rahim A., I. Y. R. Pratiwi and M. A. Fikri, “Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Classifier,” *Indonesian Journal of Computer Science (IJCS)*, vol. 12, no. 5, pp. 2995-3011, 2023.
- [3] M. F. R. Aditya, N. L. Azizah and U. Indahyanti, “Prediksi Penyakit Hipertensi Menggunakan metode Decision Tree dan Random Forest,” *Jurnal Ilmiah KOMPUTASI STI&K (Jikstik)*, vol. 23, no. 1, pp. 9-16, 2024.
- [4] M. F. Rahman, M. I. Darmawidjaja and D. Alamsah, “KLASIFIKASI UNTUK DIAGNOSA DIABETES MENGGUNAKAN METODE BAYESIAN REGULARIZATION NEURAL NETWORK (RBBN),” *JURNAL INFORMATIKA*, vol. 11, no. 1, pp. 36-45, 2017.
- [5] E. Retnoningsih and R. Pramudita, “Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python,” *BINA INSANI ICT JOURNAL*, vol. 7, no. 2, pp. 156-165, 2020.
- [6] D. H. Depari, Y. Widiastiti and M. M. Santoni, “Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung,” *Jurnal Informatik (IFTK)*, vol. 18, no. 3, pp. 239-248, 2022.
- [7] E. and S. P. Tamba, “PREDIKSI PENYAKIT GAGAL JANTUNG DENGAN MENGGUNAKAN RANDOM FOREST,” *JUSIKOM PRIMA (Jurnal Sistem Informasi dan Ilmu Komputer Prima)*, vol. 5, no. 2, pp. 176-181, 2022.
- [8] D. “Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining,” *Journal of Applied Informatics and Computing (JAIC)*, vol. 4, no. 1, pp. 84-88, 2020.
- [9] B. I. Nugroho, Z. Ma'arif and Z. Arif, “Tinjauan Pustaka Sistematis: Penerapan Data Mining Metode Klasifikasi Untuk Menganalisa Penyalahgunaan Sosial Media,” *Jurnal Sistem Informasi dan Teknologi*

- Peradaban (JSITP)*, vol. 3, no. 2, pp. 46-51, 2022.
- [10] S. B. Kotsiantis, "Decision trees: a recent overview," *Artificial Intelligence Review*, vol. 39, no. 4, pp. 261-283, 2013.
- [11] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *Cogito Smart Journal*, vol. 6, no. 2, pp. 167-178, 2020.
- [12] S. Pujiono, R. Astuti and F. M. Basysyar, "IMPLEMENTASI DATA MINING UNTUK MENENTUKAN POLA PENJUALAN PRODUK MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, pp. 615-620, 2024.
- [13] M. F. Y. Herjanto and C. , "ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI SIREKAP PADA PLAY STORE MENGGUNAKAN ALGORITMA RANDOM FOREST CLASSIFER," *JITET (Jurnal Informatika dan Teknik Elektro Terapan)*, vol. 12, no. 2, pp. 1204-1210, 2024.