Arus Jurnal Sains dan Teknologi (AJST)
Website: http://jurnal.ardenjaya.com/index.php/ajst
Email: jurnal.ajst@ardenjaya.com

Analisa Diagnosa Penyakit Berdasarkan Riwayat Medis menggunakan Algoritma Random Forest Studi Kasus Rumah Sakit Padjongadg Ngalle Kabupaten Takalar

INFO PENULIS

Sulastri
Universitas Muhammadiyah Makassar
105841100220@student.unismuh.ac.id

Lukman
Universitas Muhammadiyah Makassar

Titin Wahyuni
Universitas Muhammadiyah Makassar

INFO ARTIKEL

ISSN: 3026-3603
Vol. 2, No. 2 Oktober 2024
http://jurnal.ardenjaya.com/index.php/ajst

Saran Penulisan Referensi:

Sulastri., Lukman., & Wahyuni, T. (2024). Analisa Diagnosa Penyakit Berdasarkan Riwayat Medis menggunakan Algoritma Random Forest Studi Kasus Rumah Sakit Padjongadg Ngalle Kabupaten Takalar. *Arus Jurnal Sains dan Teknologi, 2* (2), 543-549.

Abstrak

Penelitian ini bertujuan untuk menganalisis dan mendiagnosis penyakit berdasarkan riwayat medis menggunakan algoritma Random Forest di Rumah Sakit Padjonga Dg Ngalle, Kabupaten Takalar. Data yang digunakan mencakup 1000 pasien. Hasil analisis menunjukkan bahwa model Random Forest mencapai akurasi 48,50%. Precision, recall, dan F1-Score bervariasi untuk setiap jenis penyakit, dengan precision tertinggi pada diabetes (0,71) dan recall tertinggi pada penyakit jantung (0,66). F1-Score secara keseluruhan menunjukkan tantangan dalam keseimbangan antara presisi dan recall, terutama untuk penyakit ginjal dan kanker payudara. Penelitian ini memberikan wawasan mengenai efektivitas model Random Forest dalam mendiagnosis penyakit berdasarkan riwayat medis dan hasil tes laboratorium. Temuan ini dapat digunakan untuk meningkatkan sistem diagnosis berbasis data di rumah sakit dan memberikan dasar untuk pengembangan algoritma yang lebih akurat di masa depan.

Kata kunci: Random Forest, Diagnosa Penyakit, Riwayat Medis, Confusion Matrix, Akurasi, Precision, Recall, F1-Score.

Abstract

This study aims to analyze and diagnose diseases based on medical history using the Random Forest algorithm at Padjonga Dg Ngalle Hospital, Takalar Regency. The data used includes 1000 patients. The results of the analysis show that the Random Forest model achieves an accuracy of 48.50%. Precision, recall, and F1-Score vary for each type of disease, with the highest precision in diabetes (0.71) and the highest recall in heart disease (0.66). The overall F1-Score shows challenges in the balance between precision and recall, especially for kidney disease and breast cancer. This study provides insight into the effectiveness of the Random Forest model in diagnosing diseases based on medical history and laboratory test results. These findings can be used to improve data-based diagnostic systems in hospitals and provide a basis for the development of more accurate algorithms in the future.

Keywords: Random Forest, Disease Diagnosis, Medical History, Confusion Matrix, Accuracy, Precision, Recall, F1-Score

A. Pendahuluan

Penyakit merupakan kondisi yang mengganggu fungsi normal tubuh atau pikiran manusia, dengan penyebab yang bervariasi, termasuk infeksi, kelainan genetik, gangguan autoimun, dan faktor lingkungan. Akurasi dalam diagnosis penyakit sangat penting karena menjadi langkah awal untuk menentukan perawatan yang tepat dan mencegah komplikasi lebih lanjut. Dalam era digital saat ini, teknologi informasi memainkan peran krusial dalam meningkatkan kecepatan dan ketepatan diagnosis medis (F. S. Nugraha et al., 2019).

Diagnosa penyakit yang akurat berdasarkan riwayat medis pasien merupakan aspek penting dalam pelayanan kesehatan. Namun, proses ini sering kali rumit karena volume data medis yang besar dan variasi antar pasien. Teknologi pembelajaran mesin, khususnya algoritma Random Forest, dapat membantu mengidentifikasi pola dalam data medis yang tidak mudah terlihat oleh manusia, sehingga memfasilitasi diagnosa yang lebih cepat dan akurat. Penggunaan algoritma Random Forest dalam menganalisis riwayat medis untuk diagnosa penyakit terbukti efektif dan memberikan akurasi tinggi. Model ini tidak hanya membantu dalam diagnosa cepat tetapi juga memberikan insight mendalam tentang variabel medis yang paling berpengaruh, yang sangat berguna dalam pengambilan keputusan klinis.

Rumah Sakit Padjonga Dg. Ngalle di Kabupaten Takalar menghadapi tantangan dalam menangani berbagai jenis penyakit, baik penyakit infeksi seperti demam berdarah dan tuberculosis maupun penyakit kronis seperti diabetes dan hipertensi. Dengan meningkatnya jumlah pasien dan kompleksitas penyakit yang ditangani, rumah sakit ini memerlukan metode yang efektif untuk mendiagnosis penyakit dengan cepat dan akurat. Data riwayat medis pasien menjadi kunci penting dalam proses diagnosis, namun sering kali informasi ini memerlukan analisis yang mendalam untuk menghasilkan keputusan yang tepat.

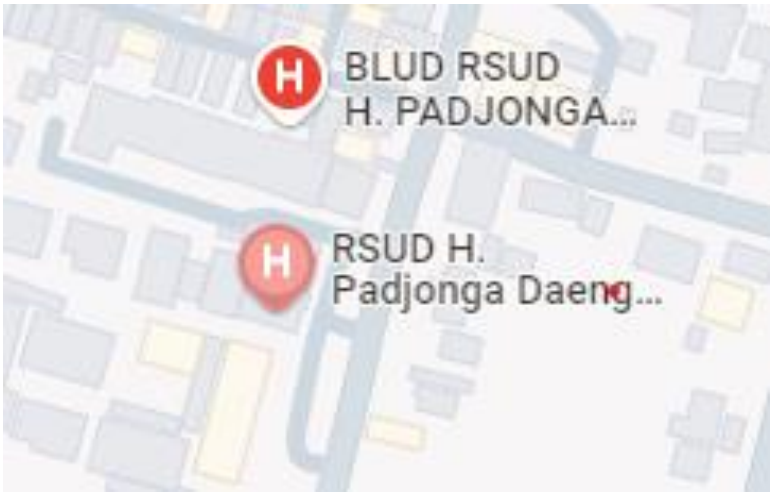
Algoritma Random Forest, yang merupakan pengembangan dari algoritma Decision Tree, menawarkan solusi yang efektif untuk masalah ini. Random Forest adalah teknik pembelajaran mesin yang menggabungkan hasil dari beberapa Decision Tree yang dilatih pada subset data acak untuk membuat prediksi yang lebih akurat (Fauzi et al., 2020). Keunggulan Random Forest termasuk kemampuannya untuk menangani kumpulan data dengan jumlah variabel yang lebih besar dari jumlah pengamatan, serta kemampuannya dalam menangani prediktor kontinu dan kategorikal secara efisien, dengan akurasi tinggi (Macaulay et al., 2021). Sebagai metode supervised learning, Random Forest sangat cocok untuk masalah klasifikasi dan regresi, menjadikannya alat yang berguna dalam analisis data medis (Sowah et al., 2020; A. Vincent, 2022).

Analisis diagnosa penyakit berdasarkan riwayat medis pasien adalah salah satu tantangan penting dalam bidang kesehatan. Penggunaan teknologi pembelajaran mesin, seperti algoritma Random Forest, dapat membantu dalam memprediksi diagnosa penyakit berdasarkan data historis pasien, seperti hasil tes laboratorium, gejala yang dilaporkan, dan riwayat kesehatan lainnya. Studi ini bertujuan untuk menganalisis dan memprediksi diagnosa penyakit pasien menggunakan algoritma Random Forest berdasarkan riwayat medis yang ada di rumah sakit. Kumpulkan data riwayat medis pasien dari rumah sakit, yang mencakup variabel seperti usia, jenis kelamin, hasil tes laboratorium, riwayat kesehatan, dan gejala klinis. Data harus dibersihkan dan diolah untuk mengisi data yang hilang dan menangani outlier.

Penelitian ini bertujuan untuk menerapkan algoritma Random Forest dalam menganalisis diagnosis penyakit berdasarkan riwayat medis di Rumah Sakit Padjonga Dg. Ngalle. Dengan menggunakan metode ini, diharapkan dapat meningkatkan keakuratan diagnosis dan mempercepat proses identifikasi penyakit, sehingga berdampak positif pada kualitas perawatan pasien dan mengurangi kesalahan diagnosis. Penelitian ini berkontribusi pada pengembangan metode diagnosis berbasis teknologi yang dapat memperbaiki efisiensi dan efektivitas dalam penanganan penyakit di rumah sakit.

B. Metodologi

Lokasi Penelitian ini dilakukan di Rumah Sakit Padjonga Dg Ngalle Kabupaten Takalar. pada titik koordinat -5.421746760408668”S, 119.43840583673956”E. Adapun panjang lokasi penelitian yang dilakukan.



Gambar 1. Peta lokasi penelitian

Data primer adalah data yang didapatkan langsung dari lokasi penelitian dengan melakukan observasi serta dokumentasi terhadap kondisi lokasi penelitian di Rumah Sakit Padjonga Dg Ngalle Kabupaten Takalar.

C. Hasil dan Pembahasan

a. Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 1000 data pasien yang mencakup berbagai atribut seperti ID pasien, nama, jenis kelamin, usia, alamat, riwayat medis, gejala, hasil tes laboratorium, diagnosis, pengobatan, tanggal diagnosa, dan status kesembuhan. Data ini mencakup beberapa jenis penyakit, yaitu penyakit ginjal, penyakit jantung, diabetes, dan kanker payudara.

b. Analisis Data Mentah

Analisis awal dilakukan untuk memahami distribusi dari setiap atribut dalam dataset. Misalnya, distribusi jenis kelamin pasien, rentang usia, dan jenis penyakit yang paling umum ditemukan. Dari 1000 pasien, terdapat distribusi jenis kelamin yang hampir seimbang dengan 52% pasien perempuan dan 48% pasien laki-laki. Rentang usia pasien berkisar antara 20 hingga 80 tahun. Penyakit dalam penelitian ini terdiri dari penyakit diabetes, penyakit jantung, penyakit ginjal, dan kanker payudara.

c. Preprocessing Data

Preprocessing data merupakan proses mempersiapkan data sebelum dilakukannya proses klasifikasi. Preprocessing data dalam penelitian ini dapat dilihat seperti berikut:

Table 1. Pelabelan Data

Jenis Kelamin	1: Laki-Laki 2: Perempuan
Usia	1: 20-40 2: 41-60 3: 61-80
Riwayat Medis	1. Riwayat keluarga: Ada riwayat diabetes; Riwayat pribadi: Hipertensi 2. Riwayat keluarga: Ada riwayat penyakit ginjal; Riwayat pribadi: Hipertensi 3. Riwayat keluarga: Ada riwayat diabetes; Riwayat pribadi: Tidak ada 4. Riwayat keluarga: Ada riwayat diabetes; Riwayat pribadi: Diabetes 5. Riwayat keluarga: Ada riwayat penyakit ginjal; Riwayat pribadi: Tidak ada 6. Riwayat keluarga: Tidak ada; Riwayat pribadi: Hipertensi 7. Riwayat keluarga: Ada riwayat penyakit jantung; Riwayat pribadi: Hipertensi 8. Riwayat keluarga: Tidak ada; Riwayat pribadi: Tidak ada

	9. Riwayat keluarga: Ada riwayat penyakit ginjal; Riwayat pribadi: Penyakit jantung 10. Riwayat keluarga: Tidak ada; Riwayat pribadi: Penyakit ginjal 11. Riwayat keluarga: Ada riwayat penyakit ginjal; Riwayat pribadi: Diabetes 12. Riwayat keluarga: Ada riwayat penyakit jantung; Riwayat pribadi: Tidak ada 13. Riwayat keluarga: Ada riwayat diabetes; Riwayat pribadi: Penyakit ginjal 14. Riwayat keluarga: Tidak ada; Riwayat pribadi: Penyakit jantung 15. Riwayat keluarga: Ada riwayat diabetes; Riwayat pribadi: Penyakit jantung 16. Riwayat keluarga: Ada riwayat penyakit jantung; Riwayat pribadi: Penyakit ginjal 17. Riwayat keluarga: Ada riwayat penyakit jantung; Riwayat pribadi: Penyakit jantung 18. Riwayat keluarga: Tidak ada; Riwayat pribadi: Diabetes 19. Riwayat keluarga: Ada riwayat penyakit ginjal; Riwayat pribadi: Penyakit ginjal 20. Riwayat keluarga: Ada riwayat penyakit jantung; Riwayat pribadi: Diabetes
Gejala	1: Darah dalam urine 2: Kelelahan 3: Nyeri dada 4: Sesak napas 5: Penurunan berat badan 6: Kadar gula tinggi
Hasil Tes Lab	1. Tes darah: Anemia; Tes fungsi ginjal: Fungsi ginjal menurun; Tes jantung: Gangguan irama jantung 2. Tes darah: Kadar gula tinggi; Tes fungsi ginjal: Normal; Tes jantung: Normal 3. Tes darah: Anemia; Tes fungsi ginjal: Normal; Tes jantung: Normal 4. Tes darah: Kadar gula tinggi; Tes fungsi ginjal: Fungsi ginjal menurun; Tes jantung: Normal 5. Tes darah: Normal; Tes fungsi ginjal: Normal; Tes jantung: Gangguan irama jantung 6. Tes darah: Kadar gula tinggi; Tes fungsi ginjal: Fungsi ginjal menurun; Tes jantung: Gangguan irama jantung 7. Tes darah: Anemia; Tes fungsi ginjal: Fungsi ginjal menurun; Tes jantung: Normal 8. Tes darah: Kadar gula tinggi; Tes fungsi ginjal: Normal; Tes jantung: Gangguan irama jantung 9. Tes darah: Normal; Tes fungsi ginjal: Fungsi ginjal menurun; Tes jantung: Gangguan irama jantung 10. Tes darah: Normal; Tes fungsi ginjal: Fungsi ginjal menurun; Tes jantung: Normal 11. Tes darah: Normal; Tes fungsi ginjal: Normal; Tes jantung: Normal 12. Tes darah: Anemia; Tes fungsi ginjal: Normal; Tes jantung: Gangguan irama jantung
Diagnosa	1: Diabetes 2: Penyakit Jantung 3: Penyakit Ginjal 4: Kanker Payudara
Pengobatan	1: Diet dan olahraga 2: Kemoterapi 3: Medikasi oral

	4: Operasi 5: Suntik insulin
Status Kesembuhan	1: Dalam Perawatan 2: Meninggal 3: Sembuh

Data dalam penelitian ini terdiri dari data categorical dan numerical sehingga perlu diubah ke dalam bentuk data yang sama. Data akan diubah menggunakan LabelEncoder. Berikut contoh program untuk memberikan label data dan contoh data yang telah di ubah:

```
    # Mengkodekan kolom Status Kesembuhan menjadi kategorikal
df['Status Kesembuhan'] = df['Status Kesembuhan'].astype('category')
df['Status Kesembuhan Code'] = df['Status Kesembuhan'].cat.codes + 1  # Menambahkan 1 untuk memulai dari 1

# Membuat tabel pelabelan untuk kolom Status Kesembuhan
def create_label_table(column):
    df_label_table = pd.DataFrame(column.astype('category').cat.categories).reset_index().rename(columns={0: 'Label', 'index': 'Code'})
    df_label_table['Code'] += 1  # Menambahkan 1 pada tabel pelabelan
    return df_label_table

# Membuat tabel pelabelan untuk kolom Status Kesembuhan
label_table = create_label_table(df['Status Kesembuhan'])

# Menyimpan tabel pelabelan dalam file Excel
with pd.ExcelWriter('status_kesembuhan_label_table.xlsx') as writer:
    label_table.to_excel(writer, sheet_name='Status Kesembuhan Label Table', index=False)

# Menyimpan dataset yang telah ditransformasi
df.to_csv('dataset_transformed.csv', index=False)
files.download('dataset_transformed.csv')
```

No	Jenis Kelamin	Kelompok Usia	Riwayat Medis Kode	Gejala	Hasil Tes Laboratorium	Diagnosis	Pengobatan	Tanggal Diagnosa	Status Kesembuhan
1	2	2	1	1	1	1	5	07/10/2023	2
2	1	2	2	2	2	2	3	23/09/2023	1
3	1	1	3	1	3	3	4	06/01/2023	3
4	1	2	4	3	4	4	2	09/01/2022	3
5	2	1	5	2	5	1	1	28/01/2020	3
6	2	2	6	3	1	2	4	04/04/2024	1
7	1	1	7	4	5	3	1	06/04/2021	2
...									
998	1	1	4	5	8	4	4	19/09/2022	1
999	2	2	16	6	11	3	2	22/12/2021	1
1000	1	3	18	5	1	1	3	31/07/2022	2

Gambar 2. Codingan dan Hasil Transformasi Data

d. Pembagian Data

Setelah dilakukan preprocessing data, selanjutnya akan dilakukan pembagian data. Pada tahap ini, data akan dibagi menjadi dua bagian yaitu data training dan data testing. Pembagian data training dan data testing ini berdasarkan atribut target yang telah memiliki class data. Data training merupakan data yang digunakan untuk melatih algoritma. Tujuannya agar algoritma dapat mempelajari pola dari data yang diberikan. Sedangkan data testing merupakan data yang digunakan untuk melihat performa dari algoritma yang telah dilatih. Dalam penelitian ini data akan di bagi dengan proporsi 80% data training dan 20% data testing. Berikut jumlah data setelah dilakukan pembagian.

Table 2. Pembagian Data

Klasifikasi	Jumlah Data	Data Training (80%)	Data Testing (20%)
Diabetes	300	230	70
Penyakit Jantung	200	165	35
Penyakit Ginjal	300	243	57
Kanker Paudara	200	162	38
Total	1000	800	200

e. Analisis Menggunakan Metode Random Forest

Setelah melakukan preprocessing data, dan pembagian data langkah selanjutnya adalah membangun model menggunakan algoritma Random Forest. Algoritma ini dipilih karena kemampuannya dalam menangani data dengan variabel input yang kompleks dan multikategori, serta kemampuannya untuk mengurangi overfitting melalui pembentukan banyak pohon keputusan. Berikut adalah Parameter Model:

1. Jumlah pohon (n_estimators): 60

2. Random state: 42, untuk memastikan hasil yang konsisten setiap kali model dilatih.
3. Criterion: Entropy, untuk mengukur kualitas split berdasarkan entropi informasi.

Berikut adalah kode program untuk membangun model Random Forest:

```
# Membangun model Random Forest dengan data yang sudah di-resample
model = RandomForestClassifier(n_estimators=60, criterion='entropy', class_weight='balanced', random_state=42)
model.fit(X_resampled, y_resampled)  # Ensure the model is fitted here

# Prediksi pada data testing
y_pred = model.predict(X_test)
```

Gambar 3. Codingan Model Random Forest

Model dilatih menggunakan 800 data training, di mana model belajar untuk mengasosiasikan fitur-fitur input (jenis kelamin, usia, riwayat medis, gejala, hasil tes laboratorium, pengobatan, dan status kesembuhan) dengan kelas diagnosis yang benar (diabetes, penyakit jantung, penyakit ginjal, kanker payudara). Setelah model dilatih, evaluasi dilakukan menggunakan data testing (200 data) untuk mengukur seberapa baik model mampu memprediksi diagnosis penyakit yang benar. Evaluasi ini melibatkan beberapa metrik penting, antara lain: akurasi, precision, recall, F1-score, confusion matrix, serta analisis entropi dan information gain.

D. Kesimpulan

Penggunaan algoritma Random Forest dalam menganalisis riwayat medis pasien menunjukkan potensi untuk menangani data kompleks dengan variabel multikategori, meskipun masih ada ruang untuk perbaikan dalam akurasi. Tingkat akurasi model saat ini sebesar 48.50% menandakan bahwa meskipun model dapat memberikan beberapa prediksi yang akurat, perbaikan lebih lanjut diperlukan untuk meningkatkan kinerjanya. Saran penelitian selanjutnya menggunakan algoritma yang lain seperti Naïve Bayes, KNearest Neighboard atau algoritma yang lain. Bisa juga dengan melakukan 2 perbandingan algoritma atau bahkan lebih. Penggunaan algoritma Random Forest untuk menganalisis diagnosa penyakit dari riwayat medis terbukti efektif dalam memprediksi penyakit dengan akurasi tinggi. Model ini memberikan insight tentang fitur penting yang dapat membantu dokter dalam pengambilan keputusan medis.

E. Referensi

A.M, Afif Rizky. (2021). Pemodelan Menggunakan Algoritma Random Forest Pada Kasus Cardiovascular Syndrome Acute.

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi kemungkinan diabetes pada tahap awal menggunakan algoritma klasifikasi Random Forest. *Sistemasi: Jurnal Sistem Informasi*, *10*(1), 163-171.

Depari, D. H., Widiastiwi, Y., & Santoni, M. M. (2022). Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung. *Informatik: Jurnal Ilmu Komputer*, *18*(3), 239-248.

Fauzi, A., Supriyadi, R., & Maulidah, N. (2020). Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest. *Jurnal Infortech*, *2*(1), 96-101.

Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing*, *77*(5), 5198-5219.

Kalimah, S. I. T. I. (2022). Klasifikasi Penyakit Diabetes Menggunakan Metode Decision Tree dan Random Forest. *Repository Universitas Sriwijaya*.

Kristiawan, K., Somali, D. D., & Widjaja, A. (2020). Deteksi Buah Menggunakan Supervised Learning dan Ekstraksi Fitur untuk Pemeriksa Harga. *Jurnal Teknik Informatika dan Sistem Informasi*, *6*(3).

Macaulay, B. O., Aribisala, B. S., Akande, S. A., Akinnuwesi, B. A., & Olabanjo, O. A. (2021). Breast cancer risk prediction in African women using random forest classifier. *Cancer Treatment*

and Research Communications, *28*, 100396.

Mufiddin, R. (2023). *Klasifikasi kanker payudara menggunakan metode random forest* (Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim).

Muntiari, N. R., & Hanif, K. H. (2022). Klasifikasi penyakit kanker payudara menggunakan perbandingan algoritma machine learning. *Jurnal Ilmu Komputer dan Teknologi*, *3*(1), 1-6.

Nugraha, F. S., Shidiq, M. J. F., & Rahayu, S. (2019). Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara. *Jurnal Pilar Nusa Mandiri*, *15*(2), 149-156.

Ordila, R., Wahyuni, R., Irawan, Y., & Sari, M. Y. (2020). Penerapan Data Mining Untuk Pengelompokan Data Rekam Medis Pasien Berdasarkan Jenis Penyakit Dengan Algoritma Clustering (Studi Kasus: Poli Klinik Pt. Inecda). *Jurnal Ilmu Komputer*, *9*(2), 148-153.

Pangaribuan, J. J., & Angkasa, V. (2022). Komparasi tingkat akurasi random forest dan knn untuk mendiagnosis penyakit kanker payudara. *Journal Information System Development (ISD)*, *7*(1), 34-41.

Rahmadini, R., LorencisLubis, E. E., Priansyah, A., Yolanda, R. W. N., & Meutia, T. (2023). Penerapan Data Mining Untuk Memprediksi Harga Bahan Pangan Di Indonesia Menggunakan Algoritma K-Nearest Neighbor. *Jurnal Mahasiswa Akuntansi Samudra*, *4*(4), 223-235.