Jurnal Informatika Medis (J-INFORMED)
Universitas Muhammadiyah Muara Bungo
DOI: http://doi.org/10.52060/im.v2i1.2119
Vol. 2, No. 1, Juni-2024, hlm. 1-4
e-ISSN: 2987-4661

METODE RANDOM FOREST UNTUK MEMUDAHKAN KLASIFIKASI DIAGNOSIS PENYAKIT MENTAL

Agus Priyono[*1]
[1]Universitas Muhammadiyah Lamongan
[1]aguspriyono2012@gmail.com

Muhammad Shodiq[2]
[2]Universitas Muhammadiyah Lamongan
[2]shodiqmuhammad13@gmail.com

Dwi Putra Alvinsyah[3]
[3]Universitas Muhammadiyah Lamongan
[3]dwialvin89@gmail.com

Septina Alfiani Hidayah[4]
[4]Universitas Muhammadiyah Lamongan
[4]septinafifi03@gmail.com

ABSTRAK

Mental health is important in the development of every individual. A bad mentality can prevent a person from developing, making a person easily stressed, hopeless, and even attempt suicide and commit crimes. Currently there are quite a lot of case related to mental health which are caused by many factors such as economic, social and medical. Reflecting on this fact, there is a need for rapid mental health detection so that immediate intervention can be carried out. This needs to be done so that the patient's condition can improve. This research focuses on diagnosing mental illness by utilizing machine learning. The method used is random forest which in several studies has been proven to produce good accuracy. Random forest performs machine learning on the attributes contained in the dataset combined with K-Fold Cross Validation so that each patient can be evaluated. Next, a tuning process is also carried out to test the parameters contained in the method. After the tuning process was carried out, the best parameters obtained were n-estimator of 30, maximum depth of 4, minimum sample leaf of 2, and minimum sample split of 10. From the combination of these parameters, accuracy is 90.83%, recall is 90.83 %, and precision of 93.25%.

Kata kunci: *mental health, random forest, machine learning, classification*

1. PENDAHULUAN

Kesehatan mental berperan krusial dalam perkembangan individu (Kinaura and Kalifia, 2024). Mental yang baik akan membuat seseorang lebih cepat berkembang baik dalam karakter, akademik, hingga fungsi otak. Sebaliknya mental yang tidak baik dapat membuat seseorang sulit berkembang, mudah stress, hingga berdampak pada perilaku kriminal (Maulana *et al.*, 2023). Untuk mengatasi tantangan ini, diperlukan solusi yang efektif untuk mempercepat proses identifikasi masalah kesehatan mental sehingga intervensi dapat diberikan lebih cepat dan tepat. Beberapa penelitian telah dilakukan dalam beberapa tahun terakhir ini. Diantara penelitian yang telah dilakukan oleh Nurhafiyah yang telah membuat sistem pakar untuk mendeteksi kesehatan mental dengan menggunakan *Certainly Factor* dengan hasil yang baik (Nurhafiyah and Marcos, 2023). Rijal mencoba membuat hal serupa namun dengan membandingkan beberapa metode. Penelitian tersebut menyimpulkan *Random Forest* menjadi metode yang paling cepat diantara metode lainnya (Rijal, Aziz and Abasa, 2024). Hidayat juga melakukan secreening terhadap *social anxiety disorder* menggunakan *Forward Chaining* (Hidayat and Mirza, 2023).

Berkaca dari pentingnya sistem diagnosis kesehatan mental, cukup banyak penelitian yang berkaitan dengan hal tersebut, baik dari sisi psikologi, kesehatan, maupun teknologi. Pada penelitian ini, diagnosa kesehatan mental dianalisa dengan menggunakan metode *Random Forest*. Metode ini dipilih berdasarkan penelitian sebelumnya yang telah membandingkan beberapa metode *machine learning* untuk kasus sesupa (Rijal, Aziz and Abasa, 2024). Hasilnya didapatkan bahwa metode *Random Forest* mendapatkan akurasi yang baik. Kemudian dataset yang

digunakan adalah data yang terpublikasi di Harvard University yang diambil dari data pemeriksaan pasien pada suatu klinik. Diharapkan metode ini bisa mendapatkan hasil akurasi yang baik dengan menggunakan dataset tersebut sehingga dapat diterapkan sebagai alat bantu pengecekan klasifikasi penyakit mental.

2. METODE PENELITIAN

Random Forest merupakan salah satu metode yang sering digunakan dalam beberapa permasalahan, diantaranya klasifikasi dan dan regresi. Metode ini diperkenalkan oleh breiman pada tahun 2001. Metode ini merupakan hasil pengembangan dari metode serupa sebelumnya yaitu CART (*Classification and Regression Tree*) dengan menggunakan teknik *bagging* atau *bootstrap aggregation* dan pemilihan fitur secara acak (A. Rahim, Inggrid Yanuar Risca Pratiwi and Muhammad Ainul Fikri, 2023). Pada penelitian ini, metode *Random Forest* akan diterapkan pada dataset yang telah didapatkan sebelumnya. Metode penelitian yang diterapkan adalah sebagai berikut

1. Persiapan dataset

Dataset diambil dari Harvard University dengan data sebanyak 120 dengan 17 atribut yang dimiliki oleh masing-masing pasien. Dataset tersebut diambil dari data pasien suatu klinik di Amerika Serikat. Masing-masing data tersebut telah diberikan label oleh dokter dengan 4 kelompok penyakit mental yaitu depresi, bipolar tipe 1, bipolar tipe 2, dan normal.

2. Preprocessing

Pada tahap ini data dievaluasi apakah ada data yang hilang, memastikan data sudah berimbang, dan melakukan normalisasi data agar hasil pembelajaran menjadi lebih maksimal. Tahapan ini merupakan hal yang penting karena dapat mempengaruhi performa dari pembelajaran mesin.

3. Proses Training dan *Tuning*

Proses ini dilakukan dengan melakukan training terhadap dataset yang telah didapatkan. Training dilakukan dengan pembentukan pola data yang didapatkan dari pembelajaran mesin terhadap dataset yang telah diberikan. Pada tahap ini juga dilakukan *tuning* untuk penyesuaian beberapa parameter yang terdapat dalam metode *Random Forest* dengan harapan ditemukan kombinasi parameter yang dapat menghasilkan akurasi terbaik. Beberapa parameter yang disesuaikan adalah *n estimators, maximum depth, minimum sample split, dan minimum sample leaf.*

4. Evaluasi

Pada tahap ini, evaluasi ditentukan untuk melihat bagaimana algoritma bekerja. Beberapa hasil akan dianalisa, terutama komposisi yang menghasilkan akurasi terbaik. Evaluasi dilakukan dengan menggunakan *K-Fold Cross Validation* dengan perbandingan 80% untuk data training dan 20% untuk data testing agar hasil yang didapat lebih merata. Hasil yang didapat merupakan hasil rata-rata dari 5 kombinasi percobaan setelah menerapkan *K-Fold Cross Validation.*

3. HASIL DAN PEMBAHASAN

Metode *Random Forest* menjadi metode yang paling akurat pada penelitian sebelumnya (Rijal, Aziz and Abasa, 2024). Penelitian tersebut membandingkan beberapa metode diantaranya *adalah Random Forest, Decision Tree, Naïve Bayes, dan K-Nearest Neighbor*. Hasil dari penelitian tersebut menyebutkan bahwa akurasi, presisi, dan *recall* dari *Random Forest* adalah sebesar 91%. Penelitian tersebut menggunakan dataset yang terdiri dari 42 atribut. Atribut-atribut tersebut merupakan data pertanyaan yang telah dijawab oleh pasien, dan juga terdapat 1 label yang menyatakan bahwa setiap pasien mengalami gangguan mental atau tidak sesuai dengan hasil pemeriksaannya. Data tersebut kemudian dinormalisasi agar rentangnya menjadi sesuai. Selanjutnya dilakukanlah pembelajaran menggunakan sebagian dataset menggunakan metode *Random Forest* sehingga terbentuklah pola pembelajaran terhadap data tersebut yang nantinya dapat digunakan untuk evaluasi model. Sebagian dataset sisanya dievaluasi menggunakan pola dari hasil pembelajaran tersebut sehingga dapat dibandingkan apakah prediksi dari model sesuai dengan data aslinya.

Pada penelitian ini terdapat perbedaan dari sisi dataset dan tahapannya. Tahapan yang berbeda adalah ketika proses pembelajaran mesin, dilakukan pula proses tuning parameter untuk menentukan parameter mana yang menghasilkan akurasi tertinggi. Kemudian ditambahkan pula proses *K-Fold Cross Validation* pada saat pembelajaran agar sebaran hasil menjadi lebih merata. Penelitian dilakukan dengan langkah-langkah sesuai yang dijelaskan pada bab sebelumnya. Detail hasil dan pembahasan setiap poin tersebut akan dijelaskan sebagai berikut.

1. Dataset

Dataset diambil dari Harvard university yang dipublikasikan pada website. Data tersebut berasal dari data pemeriksaan pasien pada sebuah klinik. Total data berjumlah 120 baris dengan jumlah atribut sebanyak 17. Setiap data tersebut telah diberikan label oleh pakar dengan 4 buah kategori yaitu depresi, bipolar tipe 1, bipolar tipe 2, dan normal.

2. Preprocessing

Dataset yang didapatkan sudah dalam bentuk balance. Setiap kategori kesehatan mental memiliki jumlah yang sama. Terdapat 17 atribut pada masing-masing data. 14 atribut diantaranya merupakan data kategori sehingga perlu dilakukan proses encoding. Kemudian dikarenakan rentang angka yang berbeda diantara atribut-atribut tersebut, maka diperlukan normalisasi terhadap data tersebut. Metode yang digunakan dalam penelitian adalah *min max normalization* (Henderi, 2021).

3. Proses Training dan *Tuning*

Tahap ini dilakukan dengan menggunakan metode *Random Forest*, serta dilakukan penyesuaian parameter terhadap metode tersebut. Pada tahap training, *Random Forest* melakukan pembelajaran dengan mengambil 80% dataset secara acak. Dari proses tersebut, pola data terbentuk sehingga dapat dilakukan evaluasi terhadap 20% data sisanya. *Random Forest* akan membentuk beberapa *tree* dengan pola tertentu. Masing-masing *tree* memiliki kriteria sesuai dengan parameter yang telah ditentukan seperti *n estimators, maximum depth, minimum sample leaf, dan minimum sample split.* Kemudian nantinya evaluasi dilakukan dengan menerapkan 20% dataset tersebut ke dalam pola hasil pembelajaran. Pengujian dilakukan dengan membandingkan hasil prediksi yang didapat dari pola pembelajaran apakah sesuai dengan hasil yang sebenarnya.

Random Forest memiliki banyak parameter yang dapat dilakukan penyesuaian (Suryadi *et al.*, 2024). Beberapa parameter yang disesuaikan adalah *n estimator, maximum depth, minimum sample split, dan minimum sample leaf.* Kombinasi nilai dari parameter tersebut dapat membuat akurasi menjadi lebih baik. Nilai-nilai yang diatur pada beberapa parameter tersebut dicari menggunakan metode *Brute Force* pada rentang nilai tertentu. Metode *Brute Force* merupakan metode pencarian secara langsung (*straight-forward*) yang berjalan menelusuri kemungkinan yang ada (Sinaga and Nuraisana, 2021).

4. Evaluasi

Tahap evaluasi dilakukan dengan kombinasi *K-Fold Cross Validation* untuk memastikan data diuji coba secara merata. *K-Fold Cross Validation* dapat membuat model *machine learning* menjadi lebih baik (Tembusai, Mawengkang and Zarlis, 2021). Nilai rata-rata dari *K-Fold Cross Validation* inilah yang dicatat sebagai hasil akurasi. Data dibagi menjadi 5 bagian atau dengan perbandingan 80:20. 80% sebagai data training dan 20% sebagai data testing. Hasil terbaik dari penelitian ini adalah dengan akurasi 90,83%, *recall* sebesar 90,83%, dan presisi sebesar 93,25%. Hasil ini didapat dari beberapa kombinasi parameter yang terlihat pada tabel 1. Hasil pengujian dengan parameter terbaik dapat dilihat pada tabel 2.

Tabel 1. Tuning Random Forest

Parameter	Nilai
N Estimator	30
Max Depth	4
Min Sample Leaf	2
Min Sample Split	10

Hasil tuning pada tabel 1 didapatkan dengan pencarian *Brute Force* pada rentang tertentu seperti pada penelitian sebelumnya yang melakukan tuning pada kasus jantung. Pada saat pembelajaran, pembentukan pola dilakukan sesuai parameter yang ditentukan secara berurutan. Dari hasil tersebut maka dicatat bahwa akurasi terbaik adalah 90% dengan nilai parameter yang tertulis pada tabel 1.

Tabel 2. Hasil evaluasi

k	Accuracy	Recall	Presition
1	95,83	95,83	97,22
2	83,33	83,33	88,69
3	83,33	83,33	87,50
4	95,83	95,83	96,30
5	95,83	95,83	96,53

Berdasarkan catatan pada tabel 2, baik akurasi, *recall*, maupun *presition* memiliki rata-rata diatas 90%. 3 dari 5*Fold* memiliki hasil yang sangat baik yakni diatas 95%. Nilai akurasi tertinggi yang bisa didapatkan adalah 95,83%, *recall* tertinggi adalah 95,83%, dan presisi tertinggi 97,22%. Diperlukan penelitian lebih lanjut untuk terus meningkatkan performa dari metode tersebut, terutama pada percobaan ke 2 dan 3 yang meskipun hasilnya bagus namun relatif lebih rendah dari pada percobaan lainnya.

Penerapan dari dianosa ini juga nantinya dapat diterapkan pada lingkungan nyata dunia kesehatan. Pada tahap awal perlu dilakukan pembelajaran mesin dengan mengacu pada dataset awal. Dokter atau petugas yang bertugas lainnya dapat mencatat nilai hasil pemeriksaan dari setiap atribut yang dimiliki oleh pasien. Kemudian sistem akan memberikan rekomendasi sesuai dengan atribut yang dimasukkan. Rekomendasi tersebut menentukan apakah pasien tersebut normal, bipolar tipe 1, bipolar tipe 2, atau depresi. Semua tahapan tersebut dilakukan tentunya setelah pola pembelajaran telah terbentuk. Selanjutnya, dataset tambahan yang baru dimasukkan setelah pola terbentuk dapat disimpan pada sistem dan nantinya dapat dilakukan proses pembelajaran mesin kembali dengan melibatkan data baru tersebut secara berkala.

4. KESIMPULAN

Hasil penelitian menunjukkan bahwa diagnosa penyakit mental dapat dilakukan dengan metode *Random Forest*. Proses *tuning* parameter juga menambah hasil akurasi dari klasifikasi ini. Nilai akurasi terbaik yang didapatkan dengan evaluasi menggunakan *K-Fold Cross Validation* adalah sebesar 90,83%, sementara Tingkat presisi adalah sebesar 93,25%, dan *recall* sebesar 90,83%. Hasil tersebut didapat setelah melakukan *tuning* parameter dengan hasil *n estimator* sebesar 30, *maximum depth* sebesar 4, *minimum sample leaf* sebesar 2, dan *minimum sample split* sebesar 10.

5. DAFTAR PUSTAKA

A. RAHIM, A.M., INGGRID YANUAR RISCA PRATIWI AND MUHAMMAD AINUL FIKRI (2023) ‘Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier’, *Indonesian Journal of Computer Science*, 12(5), pp. 2995–3011. Available at:

https://doi.org/10.33022/ijcs.v12i5.3413.

HENDERI, H. (2021) 'Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer', *IJIIS: International Journal of Informatics and Information Systems*, 4(1), pp. 13–20. Available at: https://doi.org/10.47738/ijiis.v4i1.73.

HIDAYAT, N.A. AND MIRZA, A. (2023) 'Sistem Pakar Screening Awal Gangguan Kesehatan Mental Social Anxiety Disorder Menggunakan Metode Forward Chaining Berbasis Website', *Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia*, Volume 1,(3), pp. 1–15. Available at: https://journal.mediapublikasi.id/index.php/Biner

KINAURA, N.P. AND KALIFIA, A.D. (2024) 'Gudang Jurnal Multidisiplin Ilmu Dukungan Sosial Dan Penanganan Stres Dalam Konteks Kesehatan Mental', *Gudang Jurnal Multidisiplin Ilmu*, 2, pp. 330–332.

MAULANA, M.I. *et al.* (2023) 'Upaya Penanganan Dan Peningkatan Kesehatan Mental', *KOLONI*, 2(4), pp. 90–98.

NURHAFIYAH, I. AND MARCOS, H. (2023) 'Sistem Pakar Diagnosis Kesehatan Mental Pada Mahasiswa Universitas Amikom Purwokerto', *Komputa : Jurnal Ilmiah Komputer dan Informatika*, 12(1), pp. 49–56. Available at: https://doi.org/10.34010/komputa.v12i1.8978.

RIJAL, M., AZIZ, F. AND ABASA, S. (2024) 'Prediksi Depresi : Inovasi Terkini Dalam Kesehatan Mental Melalui Metode Machine Learning Depression Prediction : Recent Innovations in Mental Health Journal Pharmacy and Application', *Journal Pharmacy and Application of Computer Sciences*, 2(1), pp. 9–14. Available at: https://doi.org/10.59823/jopacs.v2i1.47.

SINAGA, A. AND NURAISANA, N. (2021) 'Implementasi Algoritma Brute Force Dalam Pencarian Menu Pada Aplikasi Pemesanan Coffee (Studi Kasus : Tanamera Coffee)', *Jurnal Ilmu Komputer dan Sistem Informasi*, 3(3), pp. 303–313.

SURYADI, M.K. *et al.* (2024) 'A Comparative Study of Various Hyperparameter Tuning on Random Forest Classification with SMOTE and Feature Selection Using Genetic Algorithm in Software Defect Prediction', *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, 6(2), pp. 137–147. Available at: https://doi.org/10.35882/jeeemi.v6i2.375.

TEMBUSAI, Z.R., MAWENGKANG, H. AND ZARLIS, M. (2021) 'K-Nearest Neighbor with K-Fold Cross Validation and Analytic Hierarchy Process on Data Classification', *International Journal of Advances in Data and Information Systems*, 2(1), pp. 1–8. Available at: https://doi.org/10.25008/ijadis.v2i1.1204.