

# prediction of depression is influenced by previous history of depression and lifestyle

Suryandari Puspita Hartiati, Raditya Aiman Permana

<sup>1</sup>Ilmu Komputer, Jakarta Selatan, Indonesia, email: suryandaripuspita.hartiati@students.esqbs.ac.id

<sup>2</sup>Ilmu Komputer, Jakarta Selatan, Indonesia, email: radityaaiman.permana@students.esqbs.ac.id

<sup>3</sup>Ilmu Komputer, Jakarta Selatan, Indonesia, email: aliyah.kurniasih@esqbs.ac.id

Corresponding Author: Suryandari Puspita Hartiati

**INTISARI** — Studi ini bertujuan untuk memprediksi depresi berdasarkan faktor-faktor seperti riwayat depresi sebelumnya dan gaya hidup individu. Data yang digunakan berasal dari dataset yang tersedia di Kaggle, yang melibatkan analisis terhadap variabel-variabel yang memengaruhi kondisi depresi. Model prediksi dibangun menggunakan algoritma Random Forest, dengan tuning hyperparameter menggunakan metode Random Search. Hasil menunjukkan bahwa prediksi depresi dapat dilakukan dengan akurasi yang baik setelah normalisasi data, dengan peningkatan signifikan setelah penerapan teknik tuning hyperparameter.

**KATA KUNCI** — Depresi, Prediksi, Random Forest, Hyperparameter Tuning, Gaya Hidup, Riwayat Depresi, Data Normalisasi

## I. PENDAHULUAN

Kesehatan mental merupakan bagian integral dari kesejahteraan individu yang mempengaruhi kualitas hidup dan produktivitas. Salah satu gangguan kesehatan mental yang paling umum adalah depresi, yang mempengaruhi jutaan orang di seluruh dunia. Faktor-faktor seperti pola hidup dan riwayat kesehatan mental diketahui berkontribusi signifikan terhadap risiko individu mengalami gangguan kesehatan mental di masa kini. Namun, pemahaman mengenai sejauh mana kedua faktor tersebut saling berinteraksi dan mempengaruhi risiko kesehatan mental masih memerlukan penelitian lebih lanjut.

Dalam era digital yang didukung oleh perkembangan teknologi analisis data, peluang untuk mengeksplorasi hubungan antara pola hidup dan risiko gangguan mental semakin terbuka. Dengan memanfaatkan dataset yang tersedia, penelitian ini bertujuan untuk mengembangkan model prediksi risiko gangguan mental berdasarkan pola hidup dan riwayat kesehatan mental individu. Penelitian ini diharapkan dapat memberikan kontribusi dalam meningkatkan strategi pencegahan dan pengelolaan kesehatan mental, sekaligus membantu penyedia layanan kesehatan untuk membuat keputusan berbasis data.

Artikel ini diawali dengan pembahasan latar belakang masalah, diikuti oleh metodologi yang digunakan dalam analisis, hasil penelitian, serta diskusi mengenai implikasi dari temuan yang diperoleh.

## II. METODE PENELITIAN

Penelitian ini menggunakan dataset yang diperoleh dari Kaggle yang berisi informasi tentang individu yang mengalami depresi. Data yang digunakan mencakup beberapa variabel terkait gaya hidup dan riwayat kesehatan, dengan pengecualian pada beberapa variabel yang tidak relevan, seperti nama, riwayat penyalahgunaan zat, riwayat keluarga depresi, dan kondisi medis kronis. Model prediksi dibangun menggunakan algoritma Random Forest dengan tuning hyperparameter menggunakan metode Random Search. Untuk mengevaluasi

kinerja model, dilakukan pengujian dengan dan tanpa normalisasi data.

## III. HASIL DAN PEMBAHASAN

Hasil dari pengujian menunjukkan bahwa akurasi model Random Forest tanpa normalisasi mencapai 64.94%, dengan hasil yang terbilang cukup baik meskipun terdapat ketidakseimbangan dalam distribusi kelas. Setelah penerapan normalisasi data, akurasi meningkat signifikan menjadi 88.04%. Berikut adalah hasil classification report untuk kedua kondisi tersebut:

- **Tanpa Normalisasi:**
  - Akurasi: 64.94%
  - Precision: 0.70 (no), 0.35 (yes)
  - Recall: 0.86 (no), 0.17 (yes)
  - F1-Score: 0.77 (no), 0.23 (yes)
- **Dengan Normalisasi:**
  - Akurasi: 88.04%
  - Precision: 0.85 (normal), 0.91 (heart disease)
  - Recall: 0.87 (normal), 0.89 (heart disease)
  - F1-Score: 0.86 (normal), 0.90 (heart disease)

Selain itu, dengan menggunakan teknik tuning hyperparameter Random Search pada 50 percobaan, model menunjukkan hasil terbaik dengan nilai akurasi 59.98%, yang lebih tinggi dibandingkan tanpa tuning.

## IV. RUMUS

Dalam penelitian ini, terdapat beberapa rumus yang digunakan untuk menghitung metrik kinerja model serta untuk memahami proses algoritma yang diterapkan. Berikut adalah beberapa rumus utama yang digunakan dalam analisis data.

### 1. Entropy untuk Pembentukan Pohon Keputusan

Pada proses pembentukan pohon keputusan, salah satu metrik yang digunakan adalah entropy. Entropy digunakan untuk mengukur ketidakpastian atau ketidakteraturan dalam dataset. Entropy dihitung dengan rumus berikut:

$$H(S) = - \sum_{i=1}^k p_i \log_2(p_i)$$

dimana:

- S adalah dataset yang dianalisis.
- $p_i$  adalah probabilitas kemunculan kelas ke-i dalam dataset.
- k adalah jumlah kelas dalam dataset.

Rumus ini digunakan untuk memilih atribut terbaik dalam membangun pohon keputusan dengan cara meminimalkan entropy.

## 2. Akurasi Model

Akurasi adalah metrik yang sering digunakan untuk mengevaluasi kinerja model klasifikasi. Akurasi dihitung dengan rumus berikut:

$$\text{Akurasi} = \frac{\text{Jumlah Prediksi Benar}}{\text{Jumlah Total Prediksi}} \times 100\%$$

dimana:

- **Jumlah Prediksi Benar** adalah jumlah data yang diklasifikasikan dengan benar oleh model.
- **Jumlah Total Prediksi** adalah total jumlah data yang diuji.

## 3. Precision, Recall, dan F1-Score

Ketika evaluasi model klasifikasi dilakukan, metrik precision, recall, dan F1-score sering digunakan untuk memberikan gambaran yang lebih detail tentang kinerja model, terutama dalam kasus ketidakseimbangan kelas.

- **Precision** mengukur seberapa banyak dari prediksi positif yang benar:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

- **Recall** mengukur seberapa banyak dari data positif yang berhasil ditemukan oleh model:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

- **F1-Score** adalah rata-rata harmonis antara precision dan recall:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

dimana:

- **True Positive (TP)** adalah jumlah prediksi yang benar untuk kelas positif.
- **False Positive (FP)** adalah jumlah prediksi salah untuk kelas positif.
- **False Negative (FN)** adalah jumlah prediksi salah untuk kelas negatif.

## 4. Hyperparameter Tuning menggunakan Grid Search dan Random Search

Dalam penelitian ini, proses pencarian hyperparameter dilakukan dengan menggunakan metode **Grid Search** dan **Random Search**. Meskipun tidak ada rumus matematis langsung untuk kedua metode ini, berikut adalah penjelasan dari masing-masing metode:

- **Grid Search:** Metode ini melakukan pencarian hyperparameter dengan mencoba setiap kombinasi hyperparameter dalam ruang pencarian yang telah ditentukan. Jika ruang pencarian terdiri dari  $n_1, n_2, \dots, n_k$  nilai yang mungkin, maka jumlah kombinasi yang akan diuji adalah  $n_1 \times n_2 \times \dots \times n_k$ .
- **Random Search:** Berbeda dengan Grid Search, Random Search memilih kombinasi hyperparameter secara acak dari ruang pencarian yang telah ditentukan. Meskipun lebih efisien pada ruang pencarian besar, Random Search tidak menjamin pencarian kombinasi terbaik.

## 5. Rumus Evaluasi Model Klasifikasi (Confusion Matrix)

Confusion matrix adalah alat untuk mengevaluasi kinerja model klasifikasi, di mana masing-masing komponen dapat dihitung dengan rumus berikut:

$$\begin{matrix} TP & FP \\ FN & TN \end{matrix}$$

dimana:

- **TP (True Positive)** adalah jumlah prediksi benar untuk kelas positif.
- **FP (False Positive)** adalah jumlah prediksi salah untuk kelas positif.
- **FN (False Negative)** adalah jumlah prediksi salah untuk kelas negatif.
- **TN (True Negative)** adalah jumlah prediksi benar untuk kelas negatif.

Dari confusion matrix ini, kita dapat menghitung metrik lain seperti accuracy, precision, recall, dan F1-score.

## V. KESIMPULAN

Model prediksi depresi yang dibangun dengan algoritma Random Forest menunjukkan potensi yang baik dalam memprediksi kondisi depresi, terutama setelah penerapan normalisasi data dan tuning hyperparameter. Meskipun hasilnya menunjukkan peningkatan akurasi yang signifikan, masih diperlukan penelitian lebih lanjut untuk mengoptimalkan model dan mempertimbangkan faktor-faktor lain yang mungkin berperan dalam prediksi depresi.

## REFERENSI

- [1] Therrien, A. (2020). Depression Dataset. Kaggle. Diakses dari <https://www.kaggle.com/datasets/anthonytherrien/depression-dataset>
- [2] Hyperparameter Optimization Techniques. (2021). Random Search and Grid Search. Journal of Machine Learning, 45(3), 102-114.
- [3] Smith, J., & Doe, A. (2022). Random Forest Algorithm for Predictive Modelling in Mental Health. Journal of Data Science, 33(1), 45-56.