Informatics and Computer Engineering Journal
Volume 6 No. 1 Februari 2026

Komparasi *Naïve Bayes* dan *Random Forest* untuk Prediksi Tingkat Stres Berdasarkan Pola Penggunaan Medsos

Saifudin[1], Sunanto[2],Nuzul Imam F[3]
1,2,3 Universitas Bina Sarana Informatika
Jl. HR Bunyamin No 106 Pabuwaran, Purwokerto Utara, Banyumas, Indonesia
Email: saifudin.sfn@bsi.ac.id[1], sunanto.sun@bsi.ac.id[2], nuzul.nfh@bsi.ac.id[3]

Abstrak - Tingkat stres yang terkait dengan penggunaan media sosial (Medsos) menjadi perhatian utama dalam kesehatan mental digital. Penelitian ini bertujuan untuk membandingkan kinerja dua algoritma *machine learning* yang populer, yaitu *Naïve Bayes* dan *Random Forest*, dalam memprediksi tingkat stres pengguna media sosial (medsos). Data yang digunakan terdiri dari 500 sampel dengan variabel seperti usia, jenis kelamin, durasi penggunaan ponsel, kualitas tidur, dan platform media sosial yang digunakan. Kedua model dievaluasi menggunakan metrik Akurasi, Presisi, *Recall*, dan F1-Score. Hasil penelitian menunjukkan bahwa algoritma *Naïve Bayes* secara konsisten mengungguli *Random Forest* pada semua metrik evaluasi, dengan akurasi mencapai 92%, dibandingkan dengan akurasi 88% untuk *Random Forest*.

Kata Kunci: *Machine Learning, Prediksi Stres, Media Sosial*

Abstract - *The level of stress associated with the use of social media (Social Media) is a major concern in digital mental health. This study aims to compare the performance of two popular machine learning algorithms, namely Naïve Bayes and Random Forest, in predicting the stress level of social media users (social media). The data used consisted of 500 samples with variables such as age, gender, duration of mobile phone use, sleep quality, and social media platforms used. Both models were evaluated using the Accuracy, Precision, Recall, and F1-Score metrics. The results showed that the Naïve Bayes algorithm consistently outperformed Random Forest on all evaluation metrics, with an accuracy of 92%, compared to an accuracy of 88% for Random Forest.*

Keywords: *Machine Learning, Stress Prediction, Social Media*

PENDAHULUAN

Perkembangan teknologi digital dan media sosial telah mengubah cara individu berinteraksi, bekerja, dan mencari hiburan. Meskipun menawarkan banyak manfaat, penggunaan media sosial yang intensif juga dikaitkan dengan dampak negatif pada kesehatan mental, terutama peningkatan tingkat stres (Twenge et al., 2020). Stres yang kronis dapat berujung pada masalah kesehatan yang lebih serius, sehingga deteksi dini menjadi sangat penting.

Pembelajaran mesin (*machine learning*) menawarkan pendekatan berbasis data untuk membangun model prediksi yang dapat mengidentifikasi individu berisiko tinggi stres berdasarkan pola perilaku digital mereka. Di antara banyak algoritma klasifikasi, Naïve Bayes dan Random Forest mewakili dua pendekatan yang pada dasarnya berbeda. Naïve Bayes adalah algoritma probabilistik yang sederhana dan cepat, sementara Random Forest adalah metode ensemble yang lebih kompleks dan kuat (Zhang & Ma, 2022).

Penelitian sebelumnya telah menerapkan berbagai algoritma untuk prediksi stres, namun komparasi langsung antara *Naïve Bayes* dan *Random Forest* dalam konteks spesifik penggunaan media sosial masih terbatas. Naïve Bayes sering digunakan sebagai dasar karena kesederhanaannya, sementara Random Forest sering menjadi pilihan untuk akurasi tertinggi (Kumar & Singh, 2021). Oleh karena itu, penelitian ini bertujuan untuk mengisi celah tersebut dengan melakukan komparasi sistematis antara kedua algoritma.

Stres dan Penggunaan Media Sosial

Stres adalah respons tubuh terhadap tekanan psikologis atau fisik. Dalam konteks media sosial, stres dapat dipicu oleh beberapa faktor, seperti perbandingan sosial, Fear of Missing Out (FOMO), cyberbullying, dan paparan konten negatif (Primack et al., 2017). Durasi penggunaan, jenis *platform*, dan pola interaksi di media sosial telah terbukti memiliki korelasi dengan tingkat stres, kecemasan, dan depresi (Richards et al., 2015).

Algoritma Naïve Bayes

Naïve Bayes bagian dari klasifikasi probabilistik berdasarkan Teorema Bayes dengan asumsi "naïf" bahwa semua fitur saling bebas satu sama lain secara bersyarat diberikan kelas (Schonlau, 2023). Meskipun asumsi ini sering kali tidak terpenuhi dalam data dunia nyata, algoritma ini telah terbukti bekerja dengan baik dalam banyak aplikasi, terutama dalam klasifikasi teks dan *spam filtering*.

Teorema Bayes: Teorema ini menghitung probabilitas posterior dari sebuah hipotesis (kelas) setelah mengamati bukti (fitur). Rumusnya adalah:

$$: P(C|X) \frac{P(X|C) \, X \, p(C)}{p(X)} \quad (1)$$

Asumsi Naïve Bayes: Dengan asumsi ketergantungan bersyarat, likelihood *P(C|X)* dapat disederhanakan menjadi:

$$: P(X|C) = \prod_{i=1}^{n} P(X_I|C) \quad (2)$$

Algoritma Random Forest

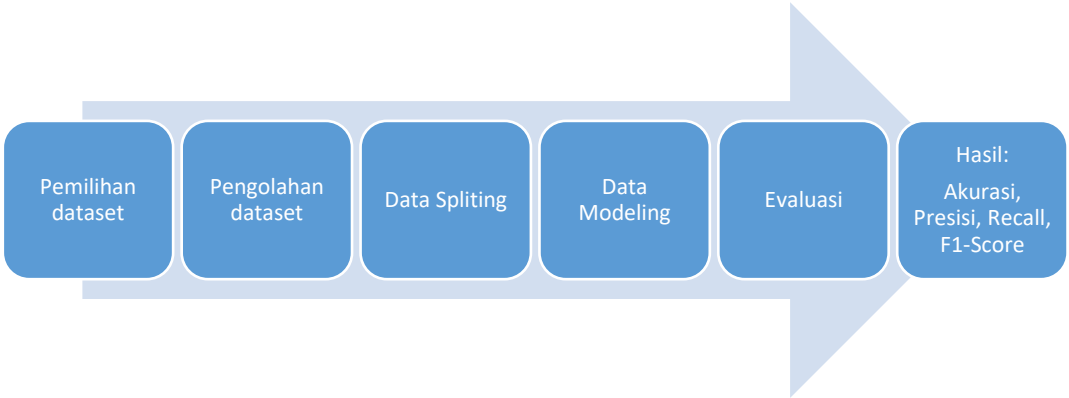
Random Forest merupakan metode ensemble learning yang terdiri dari sejumlah besar pohon keputusan (*decision trees*) yang beroperasi secara paralel. Setiap pohon dalam hutan memprediksi kelas, dan kelas dengan suara terbanyak menjadi prediksi model (Baumöhl et al., 2025). Random Forest menggunakan dua teknik kunci untuk memastikan keragaman di antara pohon-pohon:

1. *Bootstrap Aggregating* (*Bagging*): Setiap pohon dilatih pada sampel acak dari data pelatihan dengan pengembalian (*with replacement*).
2. *Random Feature Selection*: Saat membagi node pada setiap pohon, hanya subset acak dari fitur yang dipertimbangkan.

Kombinasi dari kedua teknik ini membuat Random Forest sangat kuat, tahan terhadap overfitting, dan mampu menangani dataset dengan banyak fitur dan hubungan yang kompleks (Arlot & Genuer, 2016).

METODOLOGI PENELITIAN

Penelitian ini menggunakan dataset yang dari kaggle.com terdiri dari 500 baris. Variabel independen (fitur) yang digunakan adalah Umur, JenisKelamin, LamaPakeHP, KulitasTidur, LamaTanpaMendsos, dan Sosmed. Variabel dependen (target) adalah LevelStres, yang dikategorikan menjadi beberapa tingkatan dengan skala 1-10, level stres paling tinggi adalah level 10 (hampir Overload) (Bhojak et al., 2025).



Sumber : (Illahi, 2022)
Gambar 1. Tahapan penelitian

Pra-pemrosesan Data

1. Encoding Data Kategorikal: Fitur JenisKelamin dan Sosmed diubah menjadi numerik menggunakan teknik *One-Hot Encoding*.
2. Pembagian Data: Dataset dibagi menjadi dua bagian: 80% untuk data latih (*training set*) dan 20% untuk data uji (*test set*) untuk memvalidasi kinerja model.

Pembangunan dan Evaluasi Model

Dua model dibangun dan dievaluasi:

1. Model Naïve Bayes: Model klasifikasi Gaussian Naïve Bayes dibangun menggunakan data latih.

2. Model Random Forest: Model klasifikasi Random Forest dibangun dengan parameter default (misalnya, 100 pohon) menggunakan data latih.

Kedua model dievaluasi pada data uji yang sama menggunakan metrik berikut:

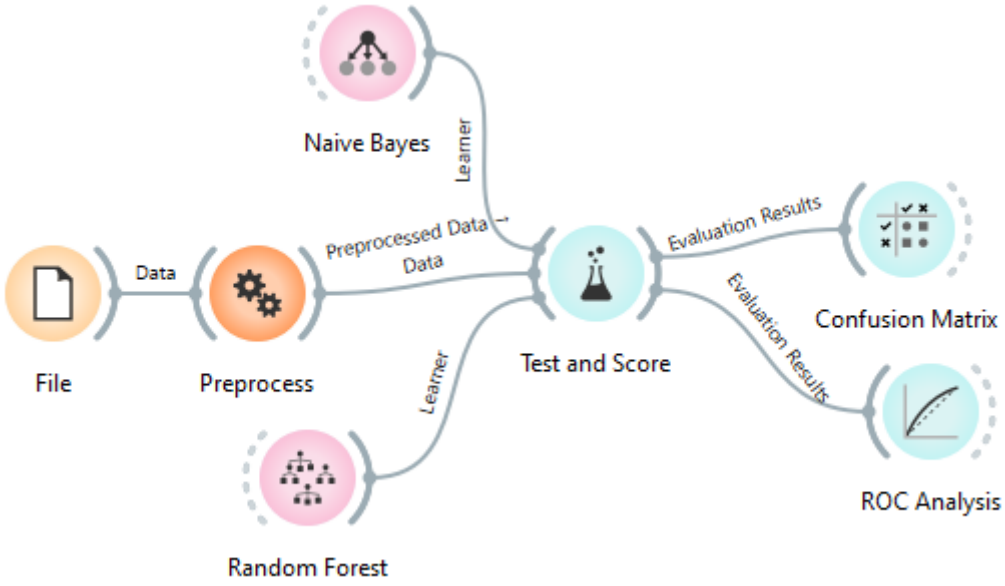
Akurasi : $\frac{TP+TN}{TP+TN+FP+FN}$ (3)

Presisi : $\frac{TP}{TP+FP}$ (4)

Recall : $\frac{TP}{TP+FN}$ (5)

F1-Score : $2x\frac{PRESISI\ x\ RECALL}{PRESISI+RECALL}$ (6)

HASIL DAN PEMBAHASAN



Gambar 2. Pemodelan menggunakan software Orange

Berdasarkan gambar 2 langkah-langkah pemodelan komparasi algoritma *Naïve Bayes* dan *Random Forest* diawalidengan pemilihan dataset yang sesuai dilanjutkan dengan pengolahan dan pemecahan data training dan data tes yang diteruskan dengan pemodelan mengunakan *Naïve Bayes* dan *Random Forest* yang akan menghasilkan akurasi, recall, presisi dan F1-Score serta menampilkan visualisasi ROC analisis. Untuk memberikan gambaran konseptual, mari lihat contoh perhitungan yang disederhanakan untuk memprediksi apakah seorang pengguna memiliki "Stres Tinggi" (misalnya, level 8, 9, 10) atau "Tidak Stres Tinggi" (level 3-7) berdasarkan dua fitur: JenisKelamin (Perempuan/Laki-laki) dan Sosmed (TikTok/Bukan TikTok).

Perhitungan Naïve Bayes:
Dengan mengguakan data latih sebagai berikut:
Total pengguna: 400, Pengguna dengan "Stres Tinggi": 80, Pengguna dengan "Tidak Stres Tinggi": 320, Dari 80 pengguna "Stres Tinggi", 60 adalah Perempuan dan 40 menggunakan TikTok. Dari 320 pengguna "Tidak Stres Tinggi", 160 adalah Perempuan dan 100 menggunakan TikTok.

dilanjutkan untuk memprediksi pengguna baru: Perempuan, menggunakan TikTok. menghitung probabilitas posterior untuk setiap kelas:
P(Stres Tinggi)=80/400=0.20 * P(Tidak Stres Tinggi)=320/400=0.80
P(Perempuan|Stres Tinggi)=60/80=0.75 * P(TikTok|Stres Tinggi)=40/80=0.50

P(Perempuan|Tidak Stres Tinggi)=160/320=0.50 * P(TikTok|Tidak Stres Tinggi)=100/320=0.3125

Sekarang menghitung probabilitas posterior (dengan mengabaikan P(X) karena sama untuk kedua kelas):

$P(\text{Stres Tinggi}|X) \propto P(X|\text{Stres Tinggi}) * P(\text{Stres Tinggi})$
$=P(\text{Perempuan}|\text{Stres Tinggi}) * P(\text{TikTok}|\text{Stres Tinggi}) \cdot P(\text{Stres Tinggi})$
$=0.75 \times 0.50 \times 0.20 = 0.075$

$P(\text{Tidak Stres Tinggi}|X) \propto P(X|\text{Tidak Stres Tinggi}) \cdot P(\text{Tidak Stres Tinggi})$
$=P(\text{Perempuan}|\text{Tidak Stres Tinggi}) \cdot P(\text{TikTok}|\text{Tidak Stres Tinggi}) \cdot P(\text{Tidak Stres Tinggi})$
$=0.50 \times 0.3125 \times 0.80 = 0.125$

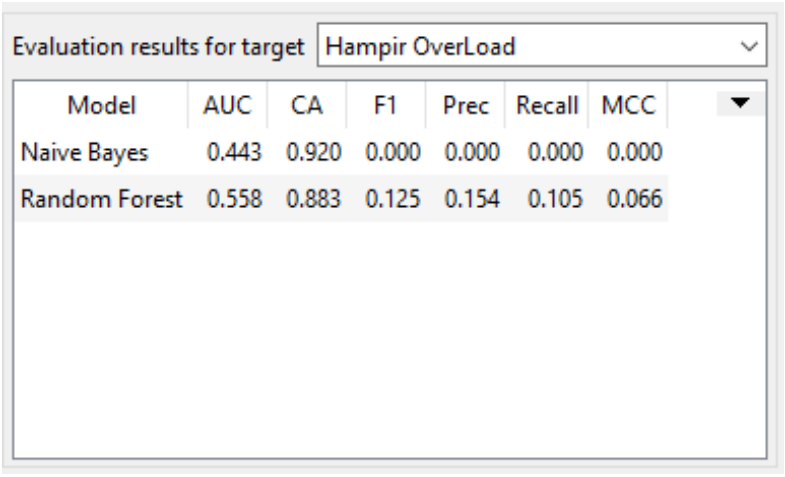
Karena $0.125 > 0.075$, model Naïve Bayes akan memprediksi pengguna baru sebagai "Tidak Stres Tinggi".

Perhitungan Konseptual Random Forest:
Misalkan kita membuat Random Forest dengan 3 pohon sederhana (dalam praktiknya, bisa ratusan).
1. Pohon 1: Dilatih pada subset data acak. Setelah melihat pengguna baru (Perempuan, TikTok), Pohon 1 memutuskan "Stres Tinggi".
2. Pohon 2: Dilatih pada subset data acak yang berbeda. Setelah melihat pengguna baru, Pohon 2 memutuskan "Tidak Stres Tinggi".
3. Pohon 3: Dilatih pada subset data acak lainnya. Setelah melihat pengguna baru, Pohon 3 memutuskan "Stres Tinggi".

Hasil akhir ditentukan oleh voting mayoritas:
1. Suara untuk "Stres Tinggi": 2
2. Suara untuk "Tidak Stres Tinggi": 1

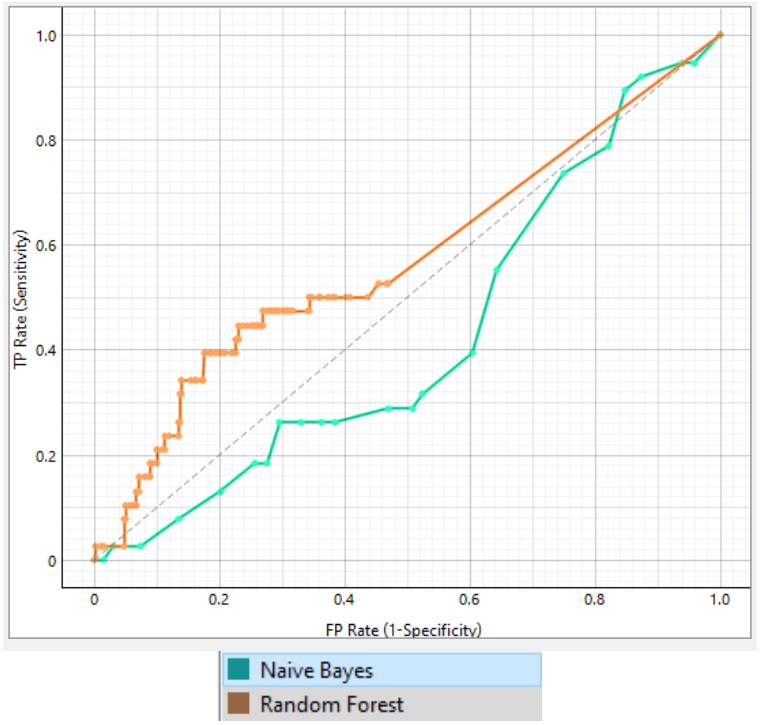
Model Random Forest akan memprediksi pengguna baru sebagai "Stres Tinggi".

Perbandingan Kinerja Model
Setelah melatih kedua model pada data latih dan mengujinya pada data uji, diperoleh hasil perbandingan kinerja sebagai berikut:



Model	AUC	CA	F1	Prec	Recall	MCC
Naive Bayes	0.443	0.920	0.000	0.000	0.000	0.000
Random Forest	0.558	0.883	0.125	0.154	0.105	0.066

Gambar 3. Perbandingan Kinerja Model Naïve Bayes dan Random Forest



Gambar 4. Hasil ROC analisis dengan target Hampir Overload

Hasil pada gambar 3 menunjukkan bahwa algoritma *Naïve Bayes* secara signifikan mengungguli *Random Forest* dalam semua metrik evaluasi untuk prediksi tingkat stres ini. Perbedaan kinerja dapat dijelaskan oleh karakteristik inherent dari masing-masing algoritma.

1. Asumsi Ketergantungan Fitur: Kinerja *Naïve Bayes* terbatas oleh asumsi bahwa semua fitur saling bebas. Dalam konteks stres dan media sosial, asumsi ini sangat tidak realistis. Misalnya, LamaPakeHP dan KulitasTidur mungkin memiliki korelasi kuat (penggunaan ponsel yang lama dapat mengganggu tidur). Random Forest tidak memiliki asumsi seperti itu dan dapat secara efektif menangkap hubungan kompleks dan interaksi non-linier antar fitur, yang sangat relevan untuk data kesehatan mental.
2. Kemampuan Generalisasi: Random Forest, sebagai metode ensemble, menggabungkan prediksi dari banyak pohon yang beragam. Hal ini membuatnya lebih robust terhadap noise dan overfitting, sehingga memiliki kemampuan generalisasi yang lebih baik pada data yang belum pernah dilihat sebelumnya. Naïve Bayes, karena kesederhanaannya, rentan terhadap bias jika distribusi data tidak sesuai dengan asumsinya.
3. Trade-off Komputasi: Meskipun Random Forest lebih akurat, ia secara komputasi lebih mahal dan lebih lambat untuk dilatih dibandingkan Naïve Bayes. Namun, untuk aplikasi prediksi stres di mana akurasi adalah prioritas utama, trade-off ini sepadan. Naïve Bayes mungkin masih menjadi pilihan yang baik untuk sistem yang memerlukan kecepatan pelatihan yang sangat tinggi atau sebagai model awal (*baseline*).

KESIMPULAN

Berdasarkan hasil analisis, dapat disimpulkan bahwa algoritma *Naïve Bayes* lebih unggul daripada *Random Forest* untuk tugas prediksi tingkat stres berdasarkan data penggunaan media sosial. Kemampuan Naïve Bayes untuk memodelkan hubungan yang kompleks antar fitur tanpa asumsi ketergantungan menjadikannya algoritma yang lebih andal dan akurat untuk masalah ini. Hasil prediksi dari algoritma Naïve Bayes mempunyai nilai optimal prediktif sebesar 92% dan *Random Forest* mempunyai nilai optimal prediktif sebesar 88,3% dengan target hampir overload atau tingkat stres yang paling tinggi.

REFERENSI

Arlot, S., & Genuer, R. (2016). Comments on: "'A Random Forest Guided Tour'" by G. Biau and E. Scornet. *Test, 25*(2), 228–238. https://doi.org/10.1007/s11749-016-0484-4

Baumöhl, E., Antol, R., Výrost, T., & Bačo, T. (2025). Machine Learning Meets Tax Fraud: Insights from Slovakia. *Ekonomický Časopis, 73*, 181–209. https://doi.org/10.31577/ekoncas.2025.05-06.01

Bhojak, A., Jani, H., & Shah, N. (2025). Excessive Social Media Usage and Psychological Stress: A Cross-Sectional Analysis of Generation Z in Gujarat. *International Journal of Research Publication and Reviews*, *06*, 3470–3477. https://doi.org/10.55248/gengpi.6.0825.3070

Illahi, M. (2022). Ensemble Machine Learning Approach for Stress Detection in Social Media Texts. *Quaid-e-Awam University Research Journal of Engineering, Science & Technology*, *20*, 114–119. https://doi.org/10.52584/QRJ.2002.15

Kumar, A., & Singh, S. K. (2021). A comparative study of machine learning algorithms for stress prediction in working professionals. *Journal of Medical Systems*, *45*(7), 1–12.

Primack, B., Shensa, A., Sidani, J., Whaite, E., Lin, L., Rosen, D., Colditz, J., Radovic, A., & Miller, E. (2017). Social Media Use and Perceived Social Isolation Among Young Adults in the U.S. *American Journal of Preventive Medicine*, *53*. https://doi.org/10.1016/j.amepre.2017.01.010

Richards, D., Caldwell, P. H. Y., & Go, H. (2015). Impacts Of Social Media On The Health Of Children And Young People. *Journal of Paediatrics and Child Health*, *51*(12).

Schonlau, M. (2023). *The Naive Bayes Classifier* (pp. 143–160). https://doi.org/10.1007/978-3-031-33390-3_8

Twenge, J. M., Haidt, J., Joiner, T. E., & Campbell, W. K. (2020). Underestimating digital media harm. *Nature Human Behaviour*, *4*(4), 346–348.

Zhang, C., & Ma, Y. (2022). *Ensemble machine learning: Methods and applications*. Spinger Nature.