TESIS

OPTIMASI ALGORITMA RANDOM FOREST UNTUK DIAGNOSIS GANGGUAN KESEHATAN MENTAL

Disusun oleh:

Nama : **Riswanto**
NIM : **22.51.2277**
Konsentrasi : **Teknologi Media Digital**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

TESIS

OPTIMASI ALGORITMA RANDOM FOREST UNTUK DIAGNOSIS GANGGUAN KESEHATAN MENTAL

OPTIMIZATION OF RANDOM FOREST ALGORITHM FOR MENTAL HEALTH DISORDER DIAGNOSIS

Diajukan untuk memenuhi salah satu syarat mencapai derajat Pascasarjana

Program Studi Informatika

disusun oleh

Nama : **Riswanto**
NIM : **22.51.2277**
Konsentrasi : **Teknologi Media Digital**

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2026

HALAMAN PERSETUJUAN

OPTIMASI ALGORITMA RANDOM FOREST UNTUK DIAGNOSIS GANGGUAN KESEHATAN MENTAL

OPTIMIZATION OF RANDOM FOREST ALGORITHM FOR MENTAL HEALTH DISORDER DIAGNOSIS

Yang disusun dan diajukan oleh

Riswanto

22.51.2277

Telah disetujui oleh Dosen Pembimbing Tesis
Pada tanggal 3 Februari 2026

Dosen Pembimbing

Tonny Hidayat, M.Kom., Ph.D.
NIK. 190302182

HALAMAN PENGESAHAN

OPTIMASI ALGORITMA RANDOM FOREST UNTUK DIAGNOSIS GANGGUAN KESEHATAN MENTAL

OPTIMIZATION OF RANDOM FOREST ALGORITHM FOR MENTAL HEALTH DISORDER DIAGNOSIS

Yang disusun dan diajukan oleh

Riswanto

22.51.2277

Telah dipertahankan di depan Dewan Penguji
pada hari Selasa, 3 Februari 2026

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Dhani Ariatmanto, S.Kom., M.Kom., Ph.D. **NIK. 190302197**	
Emha Taufiq Luthfi, S.T., M.Kom., Ph.D. **NIK. 190302125**	
Tonny Hidayat, M.Kom., Ph.D. **NIK. 190302182**	

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 3 Ferbruari 2026
DEKAN FAKULTAS ILMU KOMPUTER

Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Riswanto
NIM : 22.51.2277
Konsentrasi : Digital Transformation Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Optimasi Algoritma Random Forest Untuk Diagnosis Gangguan Kesehatan Mental
Dosen Pembimbing Utama : Tonny Hidayat, M.Kom., Ph.D.
Dosen Pembimbing Pendamping : Drs. Asro Nasiri, M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 3 Februari 2026
Yang Menyatakan,

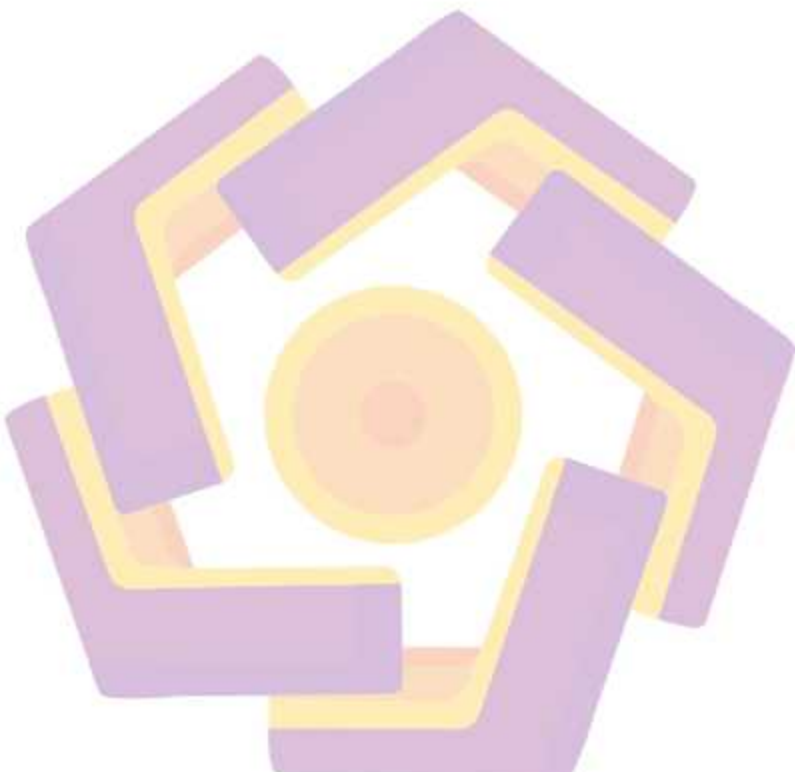
Riswanto

HALAMAN PERSEMBAHAN

Alhamdulillahrobbil'alamiin, segala puji syukur kehadirat Allah SWT yang telah melimpahkan berkah dan rahmat-Nya sehingga pada kesempatan ini saya dapat menyelesaikan tesis ini dengan baik. Saya juga mengucapkan terima kasih yang sebesar-besarnya kepada semua pihak yang telah membantu, baik secara langsung maupun tidak langsung dalam penyusunan tesis ini. Tesis ini saya dedikasikan kepada:

1. Istri terkasih Evi Widyastuti yang selalu support selama masa perkuliahan dan menyelesakan tesis sampai proses yudisium, tanpanya semua proses ini tidaklah mudah. Buah hati tersayang Aina Rasyida El Khadija yang selalu menjadi semangat tersendiri, Bapak Ibu, Mertua dan semua keluarga yang selalu mendoakan sehingga tesis ini dapat selesai dengan baik.
2. Bapak Prof. Dr. M. Suyanto, M.M. selaku rektor Universitas AMIKOM Yogyakarta yang telah memberikan kesempatan untuk dapat menimba ilmu kampus tercinta, semoga menjadi wasilah untuk kehidupan yang lebih baik di masa depan.
3. Bapak Tonny Hidayat, M.Kom., Ph.D. dan Drs. Asro Nasiri, M.Kom. selaku dosen pembimbing akademik yang selalu memberikan bimbingan dalam menyelesaikan tesis ini. Terima kasih atas ilmu yang diberikan.
4. Bapak dan Ibu Dosen yang selalu memberikan ilmu yang bermanfaat selama kuliah.
5. Terima kasih kepada teman-teman Angkatan 29 atas kebersamaan dan perjuangannya selama ini, semoga Allah mudahkan jalan kita untuk terus berkembang dan menjadi lebih baik

6. Semua pihak yang turut berkontribusi yang tidak bisa saya sebutkan satu per satu, semoga Allah limpahkan kebaikan.



HALAMAN MOTTO

Allah sesuai prasangka hambanya..

Yakinilah kebenaran dengan sepenuh hati, maka kita akan peroleh kebaikan..

Allah lebih tau kapan kita pantas memperoleh sesuatu, jadi jangan sekali kali mendekte Allah , cukup sabar dan lakukan dengan usaha terbaik yang bisa kita lakukan..

Jika ada sedikit saja terbesit rasa sombong dalam hati kita, maka yakinlah Allah pasti akan persulit jalan untuk mencapainya..

KATA PENGANTAR

Alhamdulillahirobbil'alamiin puji syukur kehadirat Allah SWT yang telah melimpahkan berkah dan rahmat-Nya sehingga penulis dapat menyelesaikan tugas akhir dengan judul **"Optimasi Algoritma Random Forest Untuk Diagnosis Gangguan Kesehatan Mental"** dengan baik dan sesuai waktu yang diharapkan.

Tesis ini disusun untuk memenuhi salah satu syarat dalam meraih gelar Magister Informatika di Universitas AMIKOM Yogyakarta. Penulis menyadari bahwa penyelesaian tesis ini tidak akan mungkin terjadi tanpa dukungan dan bantuan dari berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih yang sebesar-besarnya kepada yang terhormat:

1. Keluarga yang telah memberikan semangat, dukungan dan doanya untuk menyelesaikan studi S2 ini.
2. Bapak Prof. Dr. M. Suyanto, M.M. selaku rektor Universitas AMIKOM Yogyakarta
3. Ibu Prof. Dr. Kusrini, M.Kom selaku Direktur Program Pascasarjana.
4. Bapak Tonny Hidayat, S.Kom., M.Kom., Ph.D. dan Bapak Drs. Asro Nasiri, M.Kom. selaku dosen pembimbing akademik yang selalu memberikan bimbingan dalam menyelesaikan tesis ini.
5. Bapak Hanif Al Fatta, S.Kom., M.Kom., Ph.D. dan Bapak Emha Taufiq Luthfi, S.T., M.Kom. selaku Penguji Seminar Proposal Tesis.

6. Bapak Dhani Ariatmanto, S.Kom., M.Kom., Ph.D. dan Bapak Emha Taufiq Luthfi, S.T., M.Kom. selaku Penguji Seminar Hasil Proposal Tesis dan Pengujia Ujian Tesis
7. Kepada Bapak dan Ibu Dosen Universitas Amikom Yogyakarta yang telah memberikan ilmu, pengetahuan, motivasi, dan pengalaman selama penulis menjalani masa kuliah.
8. Teman-teman Angkatan 29 yang telah berjuang bersama selama perkuliahan dan menyelesaikan tugas akhir ini.

Penulis menyadari bahwa tesis ini masih memiliki banyak kekurangan dan kelemahan. Oleh karena itu, penulis berharap agar semua pihak dapat memberikan kritik dan saran yang membangun untuk memperbaiki dan menyempurnakan tesis ini. Penulis tetap berharap tesis ini dapat memberikan manfaat bagi semua pihak yang membacanya.

Yogyakarta, 3 Februari 2026

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Penelitian ini berjudul *"**Optimasi Algoritma Random Forest untuk Diagnosis Gangguan Kesehatan Mental**"*. Tujuan penelitian ini adalah mengklasifikasikan kondisi kesehatan mental individu di lingkungan kerja menggunakan algoritma Random Forest. Dataset yang digunakan berasal dari *Open Sourcing Mental Illness* (OSMI), sebuah organisasi non-profit yang berfokus pada isu kesehatan mental di tempat kerja. Dataset OSMI dipilih karena bersifat heterogen, terdiri dari variabel numerik dan kategorikal, sehingga sesuai dengan karakteristik algoritma *Random Forest*.

Untuk meningkatkan performa model, dilakukan optimasi *hyperparameter* menggunakan metode *Grid Search* dan *Random Search*. Selain itu, algoritma Random Forest juga dikombinasikan dengan teknik boosting, yaitu *AdaBoost* dan *XGBoost*. Beberapa skenario pengujian dilakukan guna memperoleh model dengan performa klasifikasi yang optimal.

Hasil penelitian menunjukkan bahwa hampir seluruh model memiliki performa yang baik dengan nilai akurasi di atas 80%, baik pada *Random Forest* yang dioptimasi hyperparameter maupun model yang dikombinasikan dengan teknik boosting. Model yang diusulkan juga mengalami peningkatan pada metrik evaluasi seperti akurasi, *precision, recall, F1-score*, ROC, dan AUC. Di antara seluruh model yang diuji, kombinasi *Random Forest* dengan *AdaBoost* memberikan hasil paling optimal dengan nilai akurasi sebesar 84,06% dan nilai AUC sebesar 88%, yang menunjukkan kemampuan model yang baik dalam membedakan kelas positif dan negatif.

Kesimpulan dari penelitian ini adalah bahwa integrasi *Random Forest* dengan *AdaBoost* serta optimasi hyperparameter mampu meningkatkan performa prediksi gangguan kesehatan mental di lingkungan kerja. Temuan ini diharapkan dapat membantu dalam memprediksi kondisi kesehatan mental seseorang secara lebih akurat, sehingga intervensi dan penanganan yang diberikan dapat menjadi lebih efektif dan tepat sasaran.

Kata kunci: Kesehatan Mental, OSMI, *Random Forest, Grid Search, Random Search AdaBoost, XGBoost*

ABSTRACT

This study is entitled "Optimization of the Random Forest Algorithm for Mental Health Disorder Diagnosis." The objective of this research is to classify an individual's mental health condition in the workplace using the Random Forest algorithm. The dataset used in this study was obtained from Open Sourcing Mental Illness (OSMI), a non-profit organization focused on mental health issues in the workplace. The OSMI dataset was selected due to its heterogeneous nature, consisting of both numerical and categorical variables, which aligns well with the characteristics of the Random Forest algorithm.

To improve model performance, hyperparameter optimization was conducted using Grid Search and Random Search methods. In addition, the Random Forest algorithm was combined with boosting techniques, namely AdaBoost and XGBoost. Several experimental scenarios were carried out to obtain the most optimal classification performance.

The results indicate that almost all models achieved good performance with accuracy values exceeding 80%, including both the hyperparameter-optimized Random Forest models and those combined with boosting techniques. The proposed models also demonstrated improvements across various evaluation metrics, such as accuracy, precision, recall, F1-score, ROC, and AUC. Among all evaluated models, the combination of Random Forest and AdaBoost produced the most optimal results, achieving an accuracy of 84.06% and an AUC value of 88%, indicating a strong ability to distinguish between positive and negative classes.

In conclusion, the integration of Random Forest with AdaBoost and hyperparameter optimization can significantly enhance the predictive performance of mental health disorder classification in the workplace. These findings are expected to assist in more accurate mental health predictions, enabling more effective and targeted intervention strategies.

Keywords: *Mental Health, OSMI, Random Forest, Grid Search, Random Search, AdaBoost, XGBoost.*

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

The World Health Organization (WHO) mendefinisikan kesehatan mental merupakan kondisi dimana seseorang mampu mengatasi stress dalam hidupnya sesuai dengan kemampuannya, akan tetapi tetap mampu bekerja secara normal dan produktif serta berkontribusi kepada masyarakat (Herrman et al. 2005; Yustikasari, Renata Anisa, and Retasari Dewi 2022). Pada masyarakat modern, gangguan kesehatan mental menjadi faktor penyebab berbagai penyakit kronis seperti depresi, kanker, dan penyakit kardiovaskuler yang merupakan penyebab utama kematian global (Kang et al. 2021). Dilansir dari web resmi WHO pada tahun 2019 terdapat sekitar 970 juta orang diseluruh dunia hidup dengan gangguan mental (World Health Organization 2022). Prediski WHO pada tahun 2011, bahwa pada tahun 2030 depresi menjadi sumber utama beban penyakit di seluruh dunia(Geetha et al. 2023). Sebenarnya pengobatan terhadap kesehatan mental dapat dilakukan secara efektif dengan biaya yang relatif rendah, akan tetapi sistem kesehatan masih kekurangan sumber daya yang memadai dan kesenjangan pengobatan kesehatan mental masih sangat besar di seluruh dunia (Idaiani and Riyadi 2018).

Kondisi ini menjadi keperihatinan kita bersama, sehingga diperlukan sebuah studi yang mendalam untuk mengurangi kasus gangguan kesehatan mental di masyarakat. Gangguan kesehatan mental bisa terjadi kepada siapa saja, mulai dari pelajar/mahasiswa, wanita hamil, karyawan atau bahkan orang yang sudah tua dikarenakan sakit menahun yang tak kunjung sembuh. Akan tetapi dalam penelitian ini akan penulis fokuskan pada gangguan kesehatan mental karyawan di tempat kerja. Hal ini secara tidak langsung diharapkan akan mampu mengurangi prosentase gangguan kesehatan mental karyawan di

tempat kerja karena diagnosis yang tepat. Untuk itu diperlukan metode klasifikasi yang tepat untuk membantu mendiagnosis apakah seseorang karyawan tersebut memerlukan perawatan kesehatan mental lebih lanjut ataukah tidak. Terdapat beberapa metode klasifikasi yang dapat digunakan seperti *Supporting Vector Machine* (SVM), *K-Nearset Neighbor* (KNN), Naïve Bayes, *Decision Trees*, *Random Forest*, dan *Logistic Regression* (Saputra, 2023). Dalam referensi yang lain metode klasifikasi yang bisa digunaan untuk melakukan penelitian ini antara lain *Decision Trees*, *Random Forest*, *Artificial Neural Network* (ANN), *Supporting Vector Machine* (SVM), *Linear Discriminant Analysis* (LDA), *K-Nearset Neighbor* (KNN), Logistic Regression, dan Naive Bayes (Siraj-Ud-Doulah and Alam 2018). Ada juga algoritma C4.5 seperti yang digunakan oleh (Rudi Fanani and Bendra Agustina 2024) untuk strategi penjualan produk UMKM. Terdapat pula algoritma Asosiatif yang fokus untuk menangani dataset yang besar (Nguyen et al. 2024). Pada penelitian yang lain membahas juga mengenai Ensemble Classifiers (Rezk and Selim 2024) yang kesemua algoritma klasifikasi tersebut untuk memperoleh nilai paling optimal pada setiap penelitian yang diteliti.

Berdasarkan referensi yang diperoleh, paling tidak terdapat 11 algoritma klasifikasi yang bisa digunakan yang disesuaikan dengan jenis data dan tujuan yang hendak dicapai pada setiap penelitian. Beberapa penelitian yang berkaitan dengan klasifikasi *Mental Health* telah banyak dilakukan dengan berbagai metode dan hasil klasifikasi yang berbeda. Seperti penelitian yang dilakukan oleh (Elmunsyah et al. 2019) yang melakukan klasifikasi kesehatan mental karyawan di dunia kerja dengan menggunakan metode *K-Nearest Neighbor Algorithm*. Hasil dari penelitian ini diperoleh nilai akurasi paling optimal 87,27% yang meningkat sebesar 2,27% dari penelitian

sebelumnya yang menggunakan metode Naïve Bayess dan SVM. Transformasi data kategorial menjadi bilangan biner menjadi kurang efektif karena kesehatan mental melibatkan banyak aspek yang saling berhubungan antar kategorinya. Penelitian sejenis dilakukan oleh (Sahlan et al. 2021) yang melakukan klasifikasi kesehatan mental mahasiswa menggunakan tiga algoritma *Decision Tree*, KNN dan SVM dengan nilai akurasi paling optimal pada algoritma *Decision Tree* sebesar 64%. Nilai ini sangat mungkin ditingkatkan baik dengan *hyperparaeter* maupun dengan algoritma yang lain atau menggunakan dataset yang kecukupan datanya lebih besar. Penelitian lain sejenis juga dilakukan oleh (Moningka, Hafidurrohman, and Ajri Tri 2023) yang melakukan klasifikasi kesehatan mental mahasiswa dengan menggunakan Machine Learning Naive Bayes dan KNN. Pada penelitian ini penulis nilai akurasi paling optimal adalah pada algortima KNN sebesar 85% untuk kecemasan, 80% untuk depresi, dan 70% untuk panik. Meskipun nilainya cukup baik tapi dataset yang digunakan relatif kecil yaitu 100 responden, sehingga hasil klasifikasi bisa jadi kurang representatif jika digunakan pada dataset yang lebih besar.

Selanjutnya penelitian yang dilakukan oleh (Oğur et al. 2023) yang melakukan penelitian pada masa periental menggunakan algoritma MPA dikombinasikan dengan kNN yang nilai akurasinya sangat baik mencapai 98,11% dengan menggunakan dataset primer yang berjumlah 393 data yang diperoleh dari kuesioner pada klinik rawat jalan di Rumah Sakit Penelitian Universitas Sukaraya. Hasilnya yangat sangat tinggi ini dihadapkan pada kecukupan datanya, meski dalam penelitian ini sudah cukup untuk pemodelan awal akan tetapi untuk pengembangannya perlu dilakukan penelitian dengan populasi yang lebih luas dan beragam.

Penelitian lain yang dilakukan oleh (Elisa and Rahman Isnain 2024) melakukan sentiment analisis terhadap kesehatan mental berdasarkan sosial media twitter. (Elisa and Rahman Isnain 2024) membandingakan tiga algoritma SVM, *Random Forest*, dan Naïve Bayes dengan nilai akurasi terbaik pada algoritma SVM sebesar 86,11%. Pada penelitian ini keseimbangan kelas masih menjadi permasalahan utama karena kelas negatif masih mendominasi, sehingga perlu dilakukan penanganan keseimbangan kelas agar nilai akurasi lebih optimal.

Penelitian sejenis juga dilakukan oleh (Bh et al. 2022) yang melakukan penelitian pada karyawan di tempat kerja menggunakan dataset dari website OSMI. Diperoleh nilai akurasi yang baik pada semua algoritma, dengan nilai akurasi paling optimal pada algoritma *XGBoost* sebesar 84,94%. Algoritma regresi logistik dalam penelitian ini menunjukan bahwa karyawan yang bekerja di perusahaan non-teknologi lebih banyak mengalami gangguang stress dibandingkan karyawan yang bekerja di perusahaan teknologi. Nilai ini sangat mungkin di tingkatkan dengan *hyperparameter* maupun spliting data. Dataset yang sama dignakan juga oleh (Tentua, Fidiantoro, and Ariyanto 2022) menggunakan algoritma C4.5 dan Naive Bayes. Kelemahan utama dalam penelitian ini adalah menggunakan perhitungan manual menggunakan Microsoft Excel pada pengujian probabilitas pada masing- masing klasifikasi menjadi *Yes, No*, dan *Maybe*. Selanjutnya dari 3 nilai tersebut digunakan untuk menghitung nilai probabilitas pada setiap atribut dalam dataset. Hasil akhirnya nilai *Yes* memiliki probabilitas lebih besar bila dibadingkan dengan nilai *No* dan *Maybe*. Pada penelitian ini juga tidak melakukan perhitungan matriks evaluasi sehingga nilai model yang dibuat tidak diketahui akurasinya. Penelitian lain yang dilakukan oleh (Nisa 2024) dengan dataset yang sama, menggunakan algoritma *LightGBM* diperoleh nilai akurasi paling

optimal pada *spltiting* data 90:10 dengan nilai akurasi 83%. Nilai akurasi ini sangat mungkin ditingkatkan dengan *hyperpaameter* sehingga akurasinya meningkat.

Berdasarkan beberapa penelitian di atas, maka pada penelitian ini penulis akan fokus pada klasifikasi gangguan mental di lingkungan kerja menggunakan algoritma *Random Forest*. Berdasarkan dataset dari website OSMI, permasalahan kesehatan mental di lingkungan kerja dipengaruhi oleh berbagai faktor, mulai dari aspek demografis, kondisi pekerjaan, hingga dukungan sosial dari perusahaan. Sehingga data hasil survei kesehatan mental bersifat heterogen, karena mengandung variabel numerik maupun kategorikal, serta tidak jarang terdapat nilai yang hilang maupun jawaban yang bias. Oleh karena itu, *Random Forest* dipilih karena memiliki beberapa keunggulan, di antaranya mampu mengolah data campuran dengan baik, lebih tahan terhadap *missing value* dan *noise*, serta mencegah terjadinya *overfitting* melalui teknik ensemble (Salman, Kalakech, and Steiti 2024). Selain keunggulan tersebut, *Random Forest* juga dapat memberikan interpretasi berupa *Feature Importance (Yaqoob et al. 2025)*, sehingga peneliti dapat mengetahui faktor-faktor utama yang memengaruhi kesehatan mental seorang pekerja. Selanjutnya, untuk meningkatkan nilai akurasi dari algoritma *Random Forest*, peneliti melakukan optimasi dengan menggunakan hyperparameter Grid Search dengan kombinasi dua paramaeter yang berbeda dan menggunakan hyperparameter Random Search. Keduanya dipilih karena memiliki karakteristik yang berbeda, Grid Search mencoba semua kombinasi parameter yang telah ditentukan, sedangakan Random Search memilih kombinasi parameter secara acak(Parinduri, Alkhairi, and Qurniawan 2025). Selain menggunakan *hyperparameter*, pada penelitian ini optimasi *Random Forest* juga dikombinasikan dengan teknik *boosting* yaitu AdaBoost dan XGBoost. Karakter dari dua model tersebut juga sengaja dipilih yang memiliki karakter berbeda, AdaBoost yang

cenderung cocok untuk dataset kecil sampai menengah, sedangkan XGBoost cenderung cocok untuk dataset yang besar dan kompleks (Erkamim et al. 2024). Dengan beberapa teknik optimasi tersebut, diharapkan mampu meningkatkan nilai akurasi dalam melakukan klasifikasi kesehtan mental di lingkungan kerja.

1.2. Rumusan Masalah

Latar belakang di atas menghasilkan rumusan masalah sebagai berikut

a. Berapa hasil akurasi dan apa variabel yang mempengaruhi klasfikikasi gangguan kesehatan mental menggunakan algoritma *random forest* yang dioptimasi?
b. Model optimasi *random forest* manakah yang memiliki nilai akurasi paling optimal?

1.3. Batasan Masalah

Agar pembahasan lebih fokus pada permasalahan yang sedang diteliti, maka batasan masalahnya adalah sebagai berikut :

a. Pada penelitian ini fokus pada klasifikasi gangguan kesehatan mental di dunia kerja
b. Eksperimen pada penelitian ini dilakukan dengan menggunakan platform Google Collaboratory.
c. Data yang digunakan dalam penelitian ini adalah menggunakan data yang diperoleh dari *Open Sourcing Mental Illness* (OSMI) *Mental Health in Tech Survey dataset* dengan link sebagai berikut https://osmhhelp.org/research.html
d. Jumlah dataset yang digunakan adalah sebanyak 1259 data.

e. Algoritma yang digunakan dalam penelitian ini adalah random forest yang diomtimasi menggunakan *hyperparameter Grid Search* dan *Random Search* serta dikombinasikan dengan teknik *boosting* (*AdaBoost* dan *XGBoost*)

f. Penelitian fokus untuk mengetahui performa yang paling optimal dengan menggunakan algoritma random forest yang sudah dioptimasi yang dapat dilihat dari nilai (akurasi, presisi, *recall*, *f1 score*, ROC (*Receiver Operating Characteristic*) dan AUC (*Area Under Curve*)

g. Batasan-batasan permasalahan yang akan dicari solusinya dengan penelitian yang akan dilakukan.

1.4. Tujuan Penelitian

Tujuan dari penelitan ini adalah sebagai berikut :

1.1. Untuk mengetahui klasifikasi kesehatan mental seseorang di dunia kerja sehingga memudahkan dalam pengambilan keputusan, apakah seseorang memerlukan tindakan lebih lanjut ataukah tidak.

1.2. Mengetahui nilai performa algoritma random forest yang dioptimasi menggunakan boosting (AdaBoost dan XGBoost) dengan kombnasi Random Search dan Grid Search dengan beberapa model berdasarkan evaluasi nilai akurasi, presisi, *recall*, *f1 score*, ROC dan AUC.

1.5. Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut :

a. Dapat menjadi pedoman untuk pengembangan penelitian dalam menganalisis gangguan kesehatan mental dengan berbagai studi kasus.

b. Berkontribusi secara ilmiah terhadap pengguaan algoritma random forest yang dioptimasi dalam mengklaisfikasikan gangguan kesehatan mental.

c. Hasil klasifikasi gangguan kesehatan mental diharapkan bisa menjadi bahan evaluasi bagi pihak terkait terutama rumah sakit dalam pengambilan keputusan dalam pengobatan pasien gangguan kesehatan mental.

d. Menemukan pola gangguan kesehtan mental sehingga penanganan bisa lebih cepat.

BAB II
TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Beberapa penelitian terdahulu yang pernah dilakukan dan relevan untuk dijadikan studi literatur adalah sebagai berikut :

Penulis (Sahlan et al. 2021) melakukan penelitian terkait klasifikasi kesehatan mental mahasiswa di kampus dengan menggunakan tiga algoritma yaitu *Decision Trees*, KNN dan SVM memiliki akurasi yang cukup kecil di bawah 70% pada setiap algoritma sehingga diperlukan penelitian lebih lanjut untuk meningkatkan hasil akurasi pada penelitian selanjutnya. Penelitian sejenis yang melakukan prediksi kesehatan mental di kampus yang dilakukan oleh (Moningka et al. 2023) melakukan penelitian terkait dengan klasifikasi kesehatan mental mahasiswa mengguanakan Machine Learning yang fokus pada algoritma *K-Nearest Neighbour* yang mengklasifikasikan menjadi 3 yaiut Anixety, Depresi, dan Panik dengan hasil akurasi sebesar 85% pada klasifikasi Anxiety, 80% Depresi, dan 70% untuk klasifikasi Panik.

Penelitian selanjutnya yang dilakukan oleh (Oğur et al. 2023) yang fokus pada deteksi depresi dan kecemasan pada periode periental menggunakan algoritma MPA dan kNN. Dengan kombinasi kedua metode tersebut diperoeh nilai akurasi yang cukup tinggi mencapai 98,11% dengan jumlah data yang diolah sebanyak 393 data. Penulis menyarankan untuk menambah jumlah data yang lebih banyak sehingga mampu meningkatkan model algoritma sehingga mampu diterapkan pada kelompok demografi yang lain.

Selanjutanya penelitain yang dilakukan oleh (Vaishnavi et al. 2022). Pada penelitain ini mengguankan lima (5) algoritma sekaligus dalam memprediksi kesehatan mental yaitu Logistic Regression, K-Nearest Neighbor Classifier, Decision Tree, random forests, dan Stacking dengan nilai akurasi paling tinggi menggunakan algoritma Stacking sebesar 81,75% sehingga masih sangat mungkin untuk di tingkatkan untuk hasil klasifikasi yang optimal. Pada tahun yang sama penelitain yang dilakukan oleh (Ogunseye et al. 2022) untuk memprediski kesehatan mental seseorang menggunakan algoritma AdaBoost yang bertujuan agar pengobatan lebih evektif dan efisien diperoleh nilai akurasi antara 75,93%-81,22% pada semua prediski *Machine Learning* yang artinya dengan nilai akurasi tersebut dapat digunakan untuk pengambilan keuputusan. Meskipun demikian masih sangat mungkin untuk dapat ditingkatkan sehingga prediski yang diperoleh dalam pengambilan keputusan lebih optimal.

Penelitian yang lain menggunakan algorimta KNN yang fokus pada sentiment analisis pengguna twitter untuk mendeteksi kesehatan mental seseorang melalui postingan mereka di media sosial twitter (Primadhani Tirtopangarsa and Maharani 2021). Pada penelitain ini nilai akurasi sebebsar 78,18% yang artinya masih sangat mungkin untuk dapat di tingaktkan. Penelitian yang sama dilakukan oleh (Elisa and Rahman Isnain 2024) yang melakukan sentiment analisis pada pengguna twitter diperoleh nilai akurasi maksimal pada algoritma SVM dengan nilai akurasi sebesar 86,11% diikuti algoritma random forest sebesar 82,55% dan akurasi terkecil pada algoritma Naïve Bayes dengan nilai akurasi sebesar 78,19%.

2.2. Landasan Teori

a. Klasifikasi

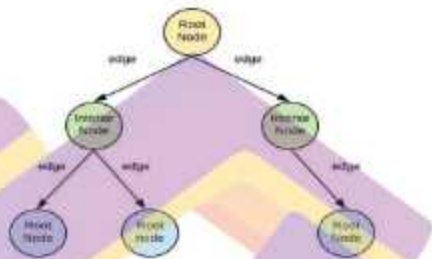
Klasifikasi secara sederhana diartikan sebagai pengelompokan, yaitu mengelompokkan sekumpulan data atau populasi berdasarkan atribut tertentu yang memiliki kemiripan satu dengan yang lainnya. Terdapat beberapa algoritma yang terkenal dalam machine learning yang memiliki model dan pendekatan yang berbeda dalam melakukan klasifiasi. Berikut adalah beberapa algoritma yang digunakan untuk klasifikasi data, yaitu : *Decision Tree, Random Forest*, KNN, SVM, dan Naïve Bayes.

Dalam penelitian ini tidak akan membahas semua jenis klasifikasi akan tetapi fokus pada algoritma Random Forest yang dioptimalkan untuk mendapatkan hasil akurasi yang lebih optimal. Harapannya dengan berbagai teknik optimasi tersebut dapat dijadikan sebagai runjukan dalam pengambilan keputusan terutama dalam kalsifikasi kesehatan mental karyawan dalam sebuah perusahaan. Beberapa teknik yang nantinya akan digunakan dalam penelitian ini antara lain : algoritma random forest dasar dengan beberapa skenario pengujian, teknik boosting dengan beberapa skenario pengujian, dan kombinasi random forest dengan teknik bosting dengan beberapa skenario pengujian yang memningkinkan diperoleh hasil paling optimal dari setiap skenario yang telah dilakukan.

b. *Decision Tree*

Decision tree merupakan salah satu algoritma klasifikasi yang menggunakan representasi struktur pohon dimna setiap node mempresentasikan atribut, cabangnya mempresentasikan nilai dari atribut, dan daun mempresentasikan kelasnya. *Decisioan tree* adalah sebuah struktur yang dapat

digunakan untuk membagi Kumpulan-kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapakan serangkaian aturan keputusan tertentu.



Gambar 2.1. Struktur *Decision Tree* (Maulida et al. 2023)

c. *Random Forest*

Random forest pertama kali diperkenalkan oleh Leo Breiman dan Adele Cutler pada tahun 2001. Algoritma ini termasuk dalam jenis algoritma *ensemble learning* yang digunakan untuk klasifikasi, regresi maupun pengelompokan data. E*nsemble learning* adalah teknik pembelajaran mesin yang menggabungkan beberapa model atau algoritma pembelajaran mesin untuk meningkatkan akurasi predisi yang dilakukan. Lebih lanjut (Saputra, 2023) menjelaskan bahwa algoritma *random forest* bekerja dengan cara menggabungkan banyak pohon keputusan (*decision tree*) yang dibangun secara acak. Misalkan $\{h(x, \theta_k), k = 1, ...\}$ dimana $\{\theta_k\}$ merupakan vector random yang iid (*independent identically distributed*) dan setiap pohon memilih kelas yang paling banyak dari data (*majority vote*). Misalkan suatu ensemble $h_1(\mathrm{x}), h_2(\mathrm{x}), ... h_k(\mathrm{x})$ dengan data

training dipilih secara random dari distribusi vector random X dan Y, fungsi margin ($mg(X,Y)$) dari random forest didefinisikan sebagai berikut :

$$mg(X,Y) = \frac{\sum_{k=1}^{K} I(h_k(X) = Y}{K} - \max_{j \neq Y} \left(\frac{\sum_{k=1}^{K} I(h_k(X) = j}{K} \right)$$

dimana *I* adalah fungsi indikasi dan *K* adalah banaknya pohon. Fungsi margin digunakan untuk mengukur Tingkat banyaknya jumlah vote pada X dan Y rata-rata vote dari kelas yang lain. *Strength* (kekuatan) adalah rata-rata (ekspektasi) ukuran kekuatan akurasi pohon tunggal. Nilai *s* yang semakin besar menunjukan bahwa akurasi prediksinya semakin baik. Nilai *s* didefinisikan sebagai berikut :

$$s = E_{X,Y}\, mg(X,Y)$$

Rata-rata korelasi (ρ) antar pasangan dugaan dari dua pohon Tunggal dalam random forest didefinisikan sebagai berikut :

$$\rho = \frac{E_{\theta,\theta'}(\rho(\theta.\theta')sd(\theta)sd(\theta'))}{E_{\theta,\theta'}(sd(\theta)sd(\theta'))}$$

dimana $\rho(\theta.\theta')sd(\theta)sd(\theta')$ merupakan korelasi antar pohon.

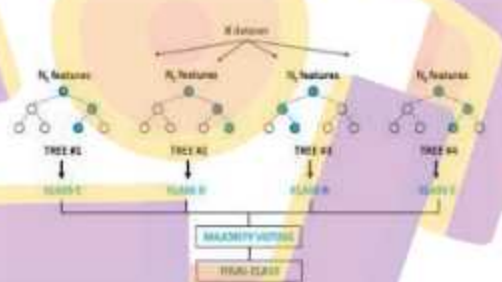
Batasan besarnya kesalahan prediksi (ε_{RF}) oleh random forest adalah :

$$\varepsilon_{RF} \leq \rho \frac{(1 - s^2)}{s^2}$$

Dari persamaan tersebut dapat dikatakan bahwa jika ingin memperoleh error yang kecil maka harus memiliki korelasi kecil dan memperkuat *strength*. Oleh karena itu diperlukan untuk mengubah nilai *m* dan *ntree*. Dengan memperkecil nilai *m*, maka memperkecil pula korelasi *strength*. Begitu pula dengan nilai *ntree*, jika nilai *ntree* besar berarti kesamaan data diantara tiap-tiap pohon sangat tinggi. Akan tetapi jika pemilihan *m* dan *ntree* sangat rendah,

mengartikan setiap pohon akan kehilangan beberapa informasi penting dan akan menaikkan nilai erornya. Sehingga pemilihan *m* dan *ntree* sangat berpengaruh terhadap hasil dari algoritma *Random Forest*.

Setiap pohon keputusan akan dihasilkan dari subset acak dari data latih. Pembentukan subset acak ini dilakukan dengan Teknik *Bagging* atau *Bootstrap Aggregating* yaitu teknik dalam *machine learning* yang digunakan untuk menghasilkan beberapa model prediksi dengan cara membangun beberapa set data latih acak dari dataset asli dan membuat model pada setiap set data latih tersebut. Hasil dari prediksi dari setiap model kemudian digabungkan menjadi satu output yang akurat dan stabil. Secara sederhana berikut Gambaran struktur dari algoritma *Random Forest*.



Gambar 2.2. Struktur *Decision Tree* (Negeri et al. 2023)

Masih dalam buku yang sama (Saputra, 2023) menguraikan beberapa hal yang berkaitan dengan algoritma *random forest*, antara lain sebagai berikut :

1. *Random forest* dapat mengatasi berbagai jenis tipe data, seeprti data numerik, kategorial, maupun kombinasi dari keduanya.
2. Mampu mengatasi *overfitting* Ketika model terlalu kompleks yang hanya berkinerja baik pada data *training*, dan berkinerja jelek pada data *testing*.

3. Algoritma *random forest* dapat menangani data yang tidak seimbang (*imbalance data*) dan tahan terhadap noise pada data.

d. Teknik Optimasi

1. Teknik Boosting

Boosting merupakan salah satu teknik optimasi dalam machine learning yaitu untuk mengurangi kesalahan dalam analisis data prediktif. Teknik Boosting ini penting karena mampu meningkatkan akurasi dan performa model yang telah dibuat dalam klasifikasi machine learning. Dalam penelitian ini untuk mencari akurasi paling optimal membandingkan dua teknik boosting yaitu AdaBoost dan XGBoost.

Adaboost merupakan merupakan singkatan dari Adaptive Boosting, adalah salah satu algoritma boosting pertama yang diperkenalkan oleh Yoav Freund dan Robert Schapire pada tahun 1996 (Saragih et al. 2021). AdaBoost adalah metode ansambel yang bertujuan untuk mengubah sejumlah model prediksi lemah (*weak learners*) yaitu model yang sedikit lebih baik dari tebakan acak yang masih rentan terhadap overfitting, menjadi model prediksi yang kuat (*strong learner*) (Febianto, Suranti, and Alinse 2024). Algoritma ini bekerja secara iteratif, di mana setiap iterasi meningkatkan performa model dengan fokus pada kesalahan yang dibuat oleh model sebelumnya (Novianti, Zarlis, and Sihombing 2022). Sehingga analisis ini akan meningkatkan akurasi sistem secara keseluruhan.

Aapun prinsip kerja dari algoritma AdaBoost dijelaskan secara rinci sebagai berikut :

a. Inisialisasi Bobot

Setiap pengamatan dalam dataset diberikan bobot yang sama di awal. Misalnya, jika terdapat N pengamatan, setiap bobot $W_1 = \frac{1}{N}$

b. Iterasi

Pada setiap iterasi t :

1. Pelatihan *Weak Learner* : Pelatihan model lemah pada data yang dibobotkan.
2. Kesalahan model : Hitung kesalahan model $h_t\ (x)$ tehadap bobot data melalui persamaan 1 di bawah ini.

$$\epsilon_t = \frac{\sum_{i=1}^{N} w_i I(y_i \neq h_t(x_i))}{\sum_{i=1}^{N} w_i}$$

dimana $I(y_i \neq h_t(x_i))$ adalah fungsi indikator yang bernilai 1 jika prediksinya salah dan bernilai 0 jika predisinya benar.

3. Bobot Model : Beri bobot pada model lemah berdasarkan kesalahanna menggunakan persamaan 2 di bawah ini.

$$\propto_i = ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

4. Pembaruan Bobot : Perbarui bobot pengamatan sehingga pengamatan yang salah klasifikasi mendapatkan bobot lebih tinggi melalui persamaan 3 di bawah ini.

$$w_i \leftarrow w_i\ exp\ (\propto_t I(y_i \neq h_t(x_i)))$$

Setelah pembaruan, bobot di-normalisasi sehingga total bobot adalah tetap 1.

c. Kombinasi Model

Model akhir adalah kombinasi dari semua model lemah yang telah diberi bobot menggunakan persamaan 4 di bawah ini.

$$H\ (x) = sign\left(\sum_{t=1}^{T} \propto_t h_t(x)\right)$$

Sedangkan XGBoost (*Extreme Gradient Boosting*) diperkenalkan oleh Tianqi Chen pada tahun 2014 (Intan Permata and Esther Sorta Mauli Nababan 2023). Algoritma ini dikembangkan sebagai bagian dari proyek *open-source* dan telah menjadi sangat populer dalam komunitas data science dan machine learning karena memiliki kinerja dan efisiensinya yang cukup tinggi. XGBoost melakukan integrasi pada beberapa teknik optimasi, yang bertujuan untuk meningkatkan kinerja dan efisiensi komputasi, termasuk regulasi, pemrosesan paralel, dan penanganan *sparsity* (Yulianti, Soesanto, and Sukmawaty 2022). XGBoost menggunakan teknik tambahan seperti shrinkage (pengurangan kontribusi setiap pohon) dan subsampling (pengambilan subset data secara acak untuk setiap pohon) untuk lebih meningkatkan kinerja (Muslim Karo Karo 2020). Teknik ini membantu mengurangi *overfitting* dan mempercepat proses pelatihan.

Prinsip kerja dari algoritma XGBoost (Extreme Gradient Boosting) dijelaskan secara rinci sebagai berikut:

a. Fungsi Tujuan

Fungsi tujuan XGBoost terdiri dari dua komponen utama: fungsi loss dan istilah regulasi untuk mengontrol kompleksitas model. Fungsi tujuan dapat dilihat pada persamaan 5 di bawah ini.

$$Obj(t) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k)$$

dimana Ω adalah istilah regulasi yang membantu mencegah terjadinya *overfitting*.

b. Istilah Regulasi

Istilah Regulasi pada XGBoost adalah regularisasi yang mengacu pada teknik yang digunakan untuk mencegah *overfitting* dengan menambahkan penalti terhadap kompleksitas model. XGBoost menggunakan dua jenis regularisasi: L1 (*Lasso*) dan L2 (*Ridge*). Regularisasi ini membantu model menjadi lebih generalis dan meningkatkan performanya pada data yang belum pernah dilihat. Regularisasi untuk kompleksitas model dihitung menggunakan persamaan 6 di bawah ini.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

dimana :

T adalah jumlah daun dalam pohon, dan

w_j adalah bobot dari daun j.

c. Fungsi Loss

Fungsi loss dihitung pada setiap iterasi melalui persamaan 7 di bawah ini.

$$Obj(t) \approx \sum_{i=1}^{n} \left[g_i f_i(x) + \frac{1}{2} h_i {f_t}^2(x_i) \right] + \Omega(f_t)$$

dimana *gi* dan *hi* adalah gradien dan hessian dari fungsi *loss*, masing-masing.

d. Pembaruan Model

Pembaruan model dilakukan dengan persamaan 8 di bawah ini.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$$

dimana η adalah laju pembelajaran.

2. Grid Search dan Random Search

Salah satu teknik dalam *machine learning* yang digunakan untuk mengoptimalkan hyperparameter tuning adalah menggunakan teknik Grid Search dan Random Search. Grid Search memiliki keunggulan pada modelnya yang sederhana dan kemampuanya untuk menemukan kombinasi hyperparameter yang paling optimal secara menyeluruh (Rahmayani and Budiman 2024). Fungi dari Grid Search adalah menguji scara istematis berbagai kemungkinan nilai parameter, sehingga menghasilkan konfigurasi terbaik bagi algoritma Random Forest dalam penelitian ini. Sehingga nilai akurasi yang dihasilkan akan lebih optimal dan kemampuan generalisasi model terhadap data baru. Berikut adalah tabel parameter yang digunakan untuk optimasi Random Forest. Menggunakan dua jenis parameter dikarenakan ada berubahan nilai akurasi meskipun tidak terlalu signifikan. Akan tetapi meskipun tidak terlalu signifikan, mampu dijadikan pertimbangan untuk dilakukan pengujian lebih lanjut supaya diperoleh nilai akurasi paling optimal dari berbagai perlakuan pada setiap parameter.

Tabel 2.1. Parameter *Grid Search* 1

Parameter	Value
n_estimators	[200, 300, 400]
max_depth	[10, 20, 30]

min_samples_split	[2, 4, 6]
min_samples_leaf	[1, 2, 4]

Tabel 2.2. Parameter *Grid Search* 2

Parameter	Value
n_estimators	[100, 200, 300]
max_depth	[10, 20, 30]
min_samples_split	[2, 4, 6]
min_samples_leaf	[1, 2, 4]

Sedangkan Random Search adalah metode optimasi hyperparameter yang akan menggabungkan elemen secara acak dalam pencarian parameter. Hal ini memngkinkan eksplorasi yang lebih luas dari ruang parameter dan membantu menemukan solusi yang efisien dalam penyetelan parameter algoritma Random Forest. Sesuai dengan namanya, metode Random Search memanfaatkan elemen secara acak dalam ekplorasi parameter, dan memilih secara acak sejumlah kombinasi parameter untuk diuji (Pramudhyta and Rohman 2024a). Berikut adalah tabel arameter untuk Random Search.

Tabel 2.3. Parameter *Random Search*

Parameter	Value
n_estimators	[100, 200, 300, 400, 500]
max_depth	[None, 10, 20, 30, 40]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4]

3. Spliting Data

Split data adalah salah satu teknik yang digunakan untuk mempartisi dataset menjadi dua bagian yaitu data latih dan data uji. Split data ini merupakan satu dari beberapa aspek yang mempengaruhi seberapa baik

kinerja model klasifikasi pada algoritma pembelajaran mesin. Metode holdout validation dan k-fold cross validation dapat digunakan untuk membagi data latih dan data uji. Proses validasi sangat penting untuk dilakukan, tujuannya agar setiap data memiliki peluang sebagai pelatihan data dan pengujian data. Spliting data merupakan salah satu teknik optimasi yang dapat mempengaruhi hasil akurasi pada suatu klasifikasi machine learning (Oktafiani, Hermawan, and Avianto 2023).

Pemilihan data latih dan data uji yang tempat sangat berpengaruh terhadap hasil akurasi yang akan dihasilkan oleh pembelajaran mesin. Beberapa penelitian membagi data latih dan data uji menjadi beberapa komposisi di antaranya yang umum digunakan adalah 90:10, 80:20, 70:30, 60:40, dan 50:50. Dari pembagian data latih ini nantinya akan diperoleh nilai akurasi paling optimal yang dapat dijadikan sebagai rujukan dalam pengambilan keputusan.

e. *Confusion Matrix*

Confusion Matrix adalah satu istilah mendasar dalam machine learning yang digunakan untuk mengukur akurasi model dengan cara membandingkan nilai prediksi dan nilai aktual (Zeng 2020). Berikut ini tabel *Confusion Matrix* menurut (Britanthia et al. 2020)

Tabel 2.4. Tabel *Conussion Matrix*

Kelas	Aktual Positif	Negatif
Prediksi Positif	Positif Benar (TP)	Positif Palsu (FP)
Negatif	Negatif Palsu (FN)	Negatif Benar (TN)

Rumus untuk menghitung nilai akurasi :

$$akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

Rums untuk menghitung nilai presisi :

$$presisi = \frac{TP}{TP + FP}$$

Rumus untuk menghitung nilai recall :

$$recall = \frac{TP}{TP + FN}$$

Rumus untuk menghitung nilai specificity :

$$specificity = \frac{TN}{TN + FP}$$

Rumus untuk menghitung F! Score :

$$F1\ Score = 2\ (\frac{recall * presisi}{recall + presisi})$$

Dimana :

TP : jumlah kelas positif yang diklasifikasi sebagai positif

FP : jumlah kelas negatif yang diklasifikasi sebagai positif

TN : jumlah kelas negative yang diklasifikasi sebagai negatif

FN : jumlah kelas positif yang diklasifikasi sebagai negatif

f. *K-fold cross-validation*

K-fold cross-validation merupakan salah satu metode validasi yang paling banyak digunakan dalam pembelajaran mesin untuk mengukur performa model prediktif. Teknik ini bekerja dengan cara membagi dataset ke dalam *K* lipatan (*fold*) yang berukuran sama, kemudian secara bergantian menggunakan satu lipatan sebagai data uji dan sisanya sebagai data latih. Proses ini diulang sebanyak *K* kali sehingga setiap data memiliki kesempatan untuk menjadi data uji maupun data latih. Pendekatan ini dinilai lebih stabil dan reliabel dibandingkan metode hold-out

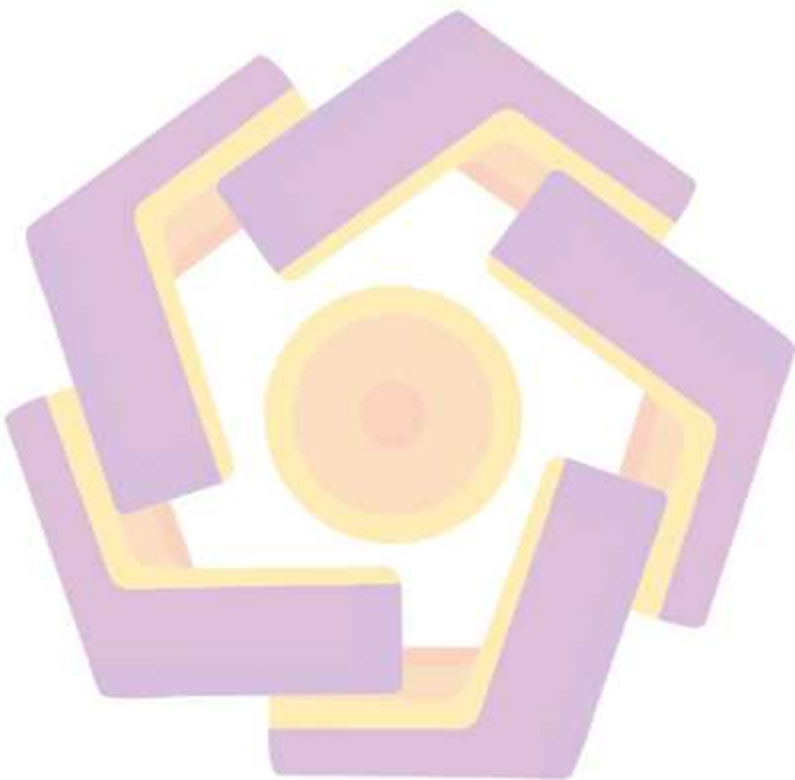
sederhana, karena memanfaatkan seluruh data yang tersedia untuk evaluasi (Li 2023).

Penelitian terbaru menunjukkan bahwa *k-fold cross-validation* tidak hanya berfungsi sebagai metode evaluasi, tetapi juga dapat dijadikan sarana inferensi statistik yang valid. (Li 2023) menemukan bahwa estimasi error yang dihasilkan oleh *k-fold cross-validation* konsisten dengan *out-of-sample error*, sehingga dapat digunakan untuk membangun interval kepercayaan performa model. Dengan demikian, metode ini mampu memberikan gambaran yang lebih komprehensif mengenai kemampuan generalisasi suatu algoritma.

Namun, penggunaan *k-fold cros-validation* juga memiliki keterbatasan bahwa, pada data dengan autokorelasi tinggi, seperti sinyal EEG, pembagian data secara acak dalam *k-fold cross-validation* dapat menyebabkan estimasi akurasi yang terlalu optimistik. Hal ini terjadi karena adanya keterkaitan antara data latih dan data uji. Untuk mengatasi hal tersebut, disarankan penggunaan variasi seperti *block-wise cross-validation* guna menjaga independensi data (White and Power 2023).

Lebih lanjut, inovasi terhadap *k-fold cross-validation* terus dikembangkan untuk meningkatkan efisiensi. (Mahlich, Vente, and Beel 2024) mengusulkan metode *e-fold cross-validation* yang mampu mengurangi konsumsi komputasi hingga 40% dengan cara menghentikan proses iterasi ketika hasil sudah stabil, tanpa menurunkan akurasi secara signifikan. Hal ini menegaskan bahwa meskipun *k-fold cross-validation* merupakan metode klasik, pengembangan variasi baru tetap

diperlukan agar validasi model dapat dilakukan dengan lebih efisien dan tetap akurat, khususnya pada dataset besar.



2.3. Keaslian Penelitian

Tabel 2.5. Matriks literatur review dan posisi penelitian
Optimasi Algoritma Random Forest Untuk Diagnosis Gangguan Kesehatan Mental

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Sama atau Kelemahan	Perbandingan
1	Detection of depression and anxiety in the perinatal period using Marine Predators Algorithm and kNN	Nur Banu Oğur, Muhammed Kotan, Deniz Balta, Burcu Çarklı Yavuz, Yavuz Selim Oğur, Hilal Uslu Yuvacı, Esra Yazıcı (2023)	Tujuan dari penelitian ini adalah untuk mendeteksi kecemasan dan depresi pada periode perintal menggunakan metode MPA dan kNN	Dalam penelitian ini kombinasi MPA dan kNN memberikan hasil akurasi paling optimal bila dibandingan dengan model yang lain yaitu sebesar 98,11 % yang mampu mendukung dalam tahap diagnosis dokter.	Penelitian ini menggunakan data sebanyak 393 sehingga diperlukan populasi data yang lebih besar untuk meningkatkan generalisasi model sehingga hasil penelitian dapat diterapkan pada berbagai kelompok demografis.	Pada penelitian menggunakan dataset publik sehingga bisa dilanjutakan untuk penelitian-penelitian selanjutanya. Selain itu metode yang digunakan menggunakan random forest yang mampu menangani jumlah data yang besar dan dioptimasi dengan boosting.
2	Mental Health Analysis of Employees using Machine Learning Techniques	Sujal BH, Neelima K, Deepanjali C, Bhuvanashree P, Kavitha Duraipandian, Sharanya Rajan, Mithileysh Sathiyanarayanan, (2022)	Tujuan dari penelitian ini adalah untuk melakukan analisis kesehatan mental karyawan menggunakan *machine learning*	Pada penelitian ini menggunakan banyak algoritma dari machine learning dalam mengukur kondisi kesehatan mental karyawan dengan nilai akurasi antara 76,34% - 84,95%.	Pada penelitian ini menggunakan banyak algoritma sehingga berpengaruh terhadap nilai akurasi yang kurang optimal. Sehingga untuk ke depan bisa dilakukan dengan algoritma tertentu dan fokus pada preproseccing data sehingga nilai akurasi lebih meningkat.	Pada penelitian ini menggunakan dataset yang sama dari OSMI sama seperti yang akan digunakan dalam penelitian sebelumnya. Perbedaannya penulis akan fokus pada algoritma random forest yang dioptimasi sehingga harapannya diperoleh nilai akurasi yang meningkat.

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	Prediction of Mental Health Among University Students	Fahruddin Sahlan, Faris Hamidi, M. Zulhafizal Misrat, M. Haziq Adli, Sharyar Wani, Yonis Gulzar (2021)	Tujuan dari penelitian ini adalah untuk memprediski kesehatan mental di kalangan mahasiswa universitas berdasar data kompetensi kewirausahaan. Algoritma yang digunakan adalah Decisioan Tree, KNN dan SVM.	Pemilihan jurusan dan jenis kelamin memilii dampak yang signifikan terhadap kesejahteraan mental mahasiswa. Dari hasil pengolahan data diperoeh bahwa algoritma Decision Tree memiliki kinerja paling baik dibandingakn degan KNN dan SVM yaitu sebesar 64%	Dataset yang digunakan memiliki 16 variabel akan tetapi jumlah data masih sangat sedikit yaitu 219 data sehingga untuk penelitian ke depan bisa dilakukan denggan mengguankan jumlah data yang lebih besar untuk meningkatkan kinerja dan validitas model prediksinya.	Pada penelitian ini penulis fokus menggunakan algoritma random forest yang dioptimasi menggunakan AdaBoost dan XGBoost untuk memperoleh nilai akurasi yang lebih baik
4	Comparison Of Random Forest, Support Vector Machine And Naive Bayes Algorithms To Analyze Sentiment Towards Mental Health Stigma	Putri Elisa, Auliya Rahman Isnain (2024)	Tujuan dari penelitian ini adalah untuk melakukan analisa sentiment pengguna twitter yang berkaitan dengan mental health mengguankan perbandingan algoritma Random Forest, SVM dan Naïve Bayes	Kesimpulan dari penelitian ini adalah klasifikasi stigma pengguan twitter tentang kesehatan mental diperoleh nilai akurasi terbaik menggunakan algorima SVM yaitu sebesar 86,11%, di ikuti random forest 82,55% dan terakhir naïve bayes 78,19%	Pada penelitian ini masih sangat mungkin untuk di tingkatkan diakrenakan nilai akurasi tertinggi baru mencapai 86% pada algoritma SVM.	Fokus dari penelitian ini adalah mengambil inti sari dari penelitian sebelumnya dengan tema yang sama akan tetapi menggunakan metode dan sasaran yang berbeda
5	Klasifikasi Mental Mahasiswa	Nirwan Moningka,	Melakukan klasifikasi ntuk	Hasil dari penelitian ini bahwa gangguan mental	Kelemahan dari penelitian ini adalah	Pada penelitian ini fokus pada klasifikasi gangguan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	Menggunakan Metode Machine Learning	Raynold, M. Hafidurrohman, Wahyu Ajri Tri R, Kusrini (2023)	mengetahui kondisi mental mahasiswa selama menghadapi pembelajaran di kampus untuk meningkatkan kesejahteraan mahasiswa	berupa stress didominasi oleh responden wanita. Dari 100 responden sebanyak 94% orang yang memerlukan tindakan lebih lanjut dengan nilai akurasi tertinggi tertinggi sebesar 85% menggunakan algoritma KNN pada klasifikasi *Anxiety*	jumlah sampel yang terbatas sebanyak 100 data dengan mayoritas responden wanita sehingga belum bisa mewakili kondisi mahasiswa secara umum. Variabel yang digunakan juga masih terbatas sehingga berpengaruh terhadap hasil dari metode yang digunakan dalam penelitian ini.	kesehatan mental menggunakan random forest pada lingkungan kerja dengan dataset dari OSMI.
6	Classification of Employee Mental Health Disorder Treatment with K-Nearest Neighbor Algorithm	Hakkun Elmunsyah, Risalatul Mu'awanah, Triyanna Widiyaningtyas (2019)	Tujuan penelitian ini adalah untuk mengklasifikasikan data perawatan kesehatan mental karyawan menggunakan metode KNN	Kesimpulan dari penlitian ini adalah algoritma KNN merupakan metode yang efektif untuk klasifikasi data perawatan kesehatan mental karyawan karena hasil akurasi mencapai 87,27% dengan perbandingan data training dan data testing sebesar 85:15	Secara umum baik akan tetapi perlu ditambahkan algoritma lain sebagai pembanding sehingga bisa diketahui algoritma yang memberikan hasil paling optimal.	Pada penelitian ini penulis menggunakan algoritma random forest yang dioptimasi dengan teknik boosting sehingga harapannya akan diperoleh nilai akurasi yang optimal.

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat dan Pendekatan Penelitian

Jenis, sifat, dan pendekatan pada penelitian yang akan dilaksanakan pada penelitian ini adalah sebagai berikut :

1. Jenis penelitian

   Jenis penelitain ini adalah jenis penelitian eksperimental dimana penelitian ini melakukan pengujian tingkat akurasi tertinggi menggunakan metode random forest dibandingkan dengan random forest yang telah dioptimasi dengan mengunakan dataset yang sama. Pengujian ini bertujuan untuk mengetahui metode yang paling optimal dalam melakukan klasifikasi untuk diagnosis gangguang kesehatan mental sehingga dapat digunakan untuk dasar pengambilan tindakan pengobatan lebih lanjut ataukah tidak.

2. Sifat Penelitian

   Penelitian ini bersifat deskriptif, diakrenakan menggambarkan suatu objek yang akan diteliti dan menjabarkan hasil dari pengujian-pengujian yang dilakukan pada dataset yang telah ditentukan sebelumnya untuk dapat diketahui metode mana yang memiliki akurasi, presisi, *recall* dan *F1 score* yang paling optimal.

3. Pendekatan Penelitian

   Penelitian ini menggunakan pendekatan kualitatif karena menggunakan angka-angka dan grafik yang hasil dari ekperimen menggunakan

algoritma random forest dapat digunakan untuk penarikan Kesimpulan berdasarkan akurasi yang paling optimal.

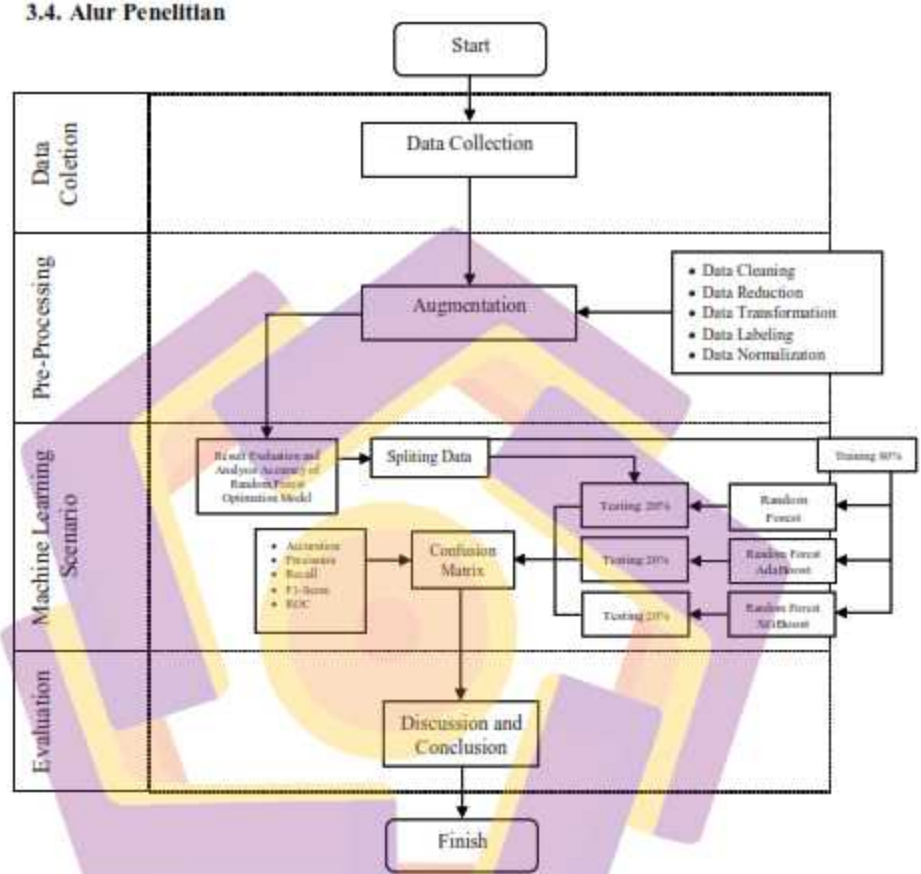
3.2. Metode Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah menggunakan data yang diperoleh dari *Open Sourcing Mental Illness* (OSMI) *Mental Health in Tech Survey dataset* dengan link sebagai berikut https://osmhhelp.org/research.html. Dari data OSMI tersebut kemudian difokuskan pada data di tahun 2016 dengan jumlah data sebanyak 1259 data. OSMI sendiri adalah salah satu lembaga non-profit yang menyadari betapa pentingnya masalah kesehatan mental pada lingkungan kerja dan bergerak dalam bidang kesehatan terutama kesehatan mental pada pekerja dalam industri teknologi (Alfaresy Chaerudin et al. 2022).

3.3. Metode Analisis Data

Sebelu melakukan pre-processing dataset yang diperoleh terlebih dahulu dilakukan pemodelan dan pelabelan data disesuaikan dengan kebutuhan. Data yang sudah diberi label tersebut kemudian dibagi menjadi dua jenis data yaitu data *training* dan data *testing*. Data tersebut kemdian dianalisis menggunakan algoritma random forest yang sudah diotimasi untuk didapatkan tingkat akurasi, presisi, *recall* dan *F1 score*. Hasil dari pengukuran yang paling optimal tersebut akan dijadikan sebagai acuan atau pedoman dalam menentukan diagnosis tindakan dalam penanganan gangguan kesehatan mental.

3.4. Alur Penelitian



Gambar 3.1 Alur Penelitian

5. Data Collection

Penelitian ini dimulai dari tahap studi literatur, pada tahap ini penelitian dimulai dengan idntifikasi masalah dengan melakukan analisis latar belakang masalah, merumuskan masalah, menentukan tujuan penelitian, dan manfaat dari penelitianyang akan dilakukan. Metode pengumpulan datanya adalah dengan cara observasi pada beberapa sumber referensi yang akhirnya diputusakan untuk menggunakan dataset dari OSMI yang dikumpulkan pada tahun 2016 dengan jumlah data sebanyak 1259 data

yang berkaitan dengan kesehatan mental karyawan pada sebuah perusahaan.

6. Pre-Processing

   Data akan melalui proses pre-processing untuk memastikan data dapat dioleh tanpa ada gangguan yang tidak perlu sehingga tidak mempengaruhi hasil perhitungan.

   Proses augmentasinya adalah dengan melakukan data cleaning, data reduction, data transformation, data labeling, dan data normalization.

7. Machine Learning Scenario

   Pada tahap ini dilakukan spliting data pada setiap skenario yang telah ditentukan yaitu Random Forest, Random Forest AdaBoost dan Random Forest XGBoost untuk memperoleh nilai akurasi paling tinggi dari setiap skenario yang telah dilakuan. Harapnya dengan nilai akurasi paling optimal dapat dijadikan sebagai rujukan dalam pengambilan keputusan berdasarkan hasil dari penelitian ini.

8. Evaluation

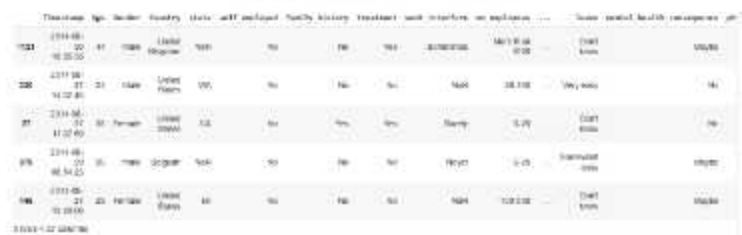
   Pada tahap ini dari setiap skenario dapat ditetnukan aspek yang paling berpengaruh terhadap hasil perhitungan sehingga bisa dijadikan sebagai argumentasi pada setiap pengambilan keputusan.

BAB IV
HASIL PENELITIAN DAN PEMBAHASAN

4.1. Pengumpulan Dataset

Proses pengumpulan dataset dilakukan dengan mencari dari beberapa referensi yang menggunakan dataset yang sama, dengan tujuan untuk memastikan dataset tersebut layak untuk digunakan pada penelitian ini. Pada akhirnya ditemuakan ada sekitar empat (4) referensi yang menggunakan dataset yang sama yaitu dari website OSMI (*Open Sourcing Mental Illness) Mental Health in Tech Survey dataset* dengan link sebagai berikut https://osmhhelp.org/research.html. Pada website tersebut terdapat beberapa dataset, kemudian dipilih data pada tahun 2016 sberdasarkan referensi yang diperoleh dengan jumlah data sebanyak 1259. Dari data 1259 tersebut memiliki sekitar 26 variabel yang dijadikan sebagai acuan untuk meneliti tingkat kesehatan mental karyawan pada sebuah perusahaan. Gamabr 4.1 di bawah ini adalah tampilan data mentah yang diperoleh langsung dari website OSMI, sedangkan gambar 4.2 adalah tampilan setelah masuk di Google Colaboratory.

Gambar 4.1 Data Mentah OSMI



Gambar 4.2 Tampilan Data Google Colaboratory

Lebih lanjut untuk mengetahui persebaran data berdasarkan beberapa variabel, ditampilkan beberapa data berdasarkan lokasi tempat bekerja, usia, kemudahan dalam penagmbilan cuti (leave), serta apakah pernah menerima perawatan atau tidak (treatment).

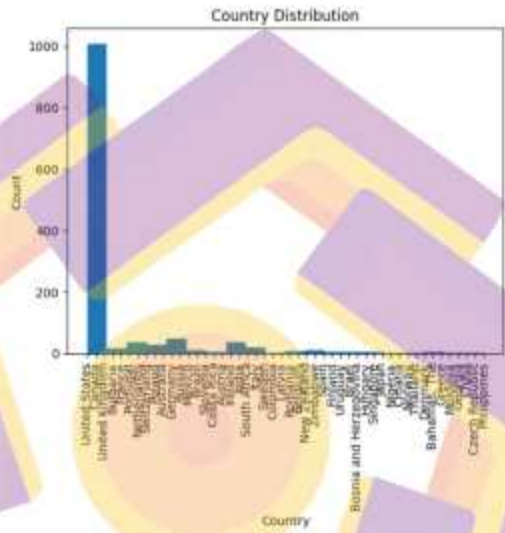


Diagram 4.1 Distribusi Negara Tempat Bekerja

Berdasarkan diagram di atas, untuk responden dari data yang digunakan mayoritas berada di Amerika Serikat sekitar 1000 data kemudian di ikuti negara-negara lain seperti Canada, Germany, United Kingdom yang jumlah responden dari setiap negara tidak lebih dari 100 orang.

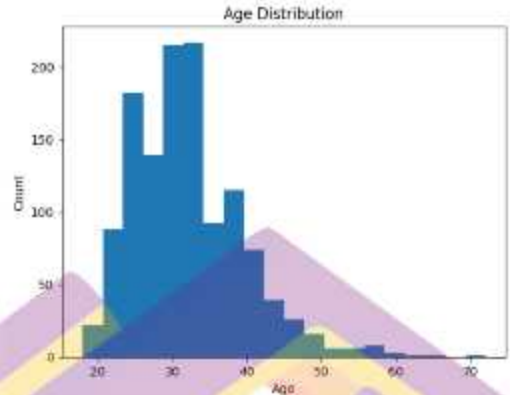


Diagram 4.2 Distribusi Usia Pekerja

Berdasarkan diagram di atas, responden didominasi umur antara 20 tahun sampai dengan 70 tahun dengan responden terbanyak pekerja usia produktif antara umur 25-40 tahun.

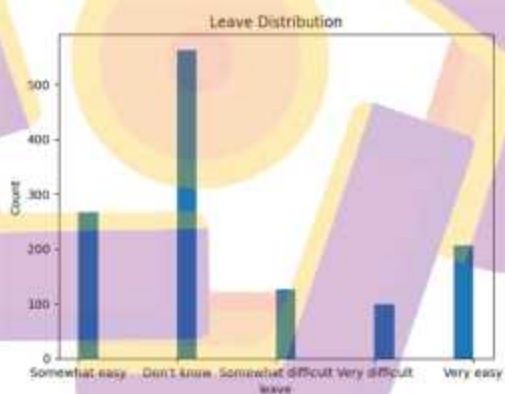


Diagram 4.3 Distribusi Kemudahan Cuti

Pada diagram di atas rata-rata pekerja tidak mengetahui apakah mereka dapat mengajukan cuti dengan mudah. Sedangkan dibeberapa perusahaan masih ada yang kesulitan dalam mengajukan cuti meskipun jumlahnya kecil akan tetapi tetap signifikan dalam mempengaruhi kondisi kesehatan mental karyawan di tempat kerja.

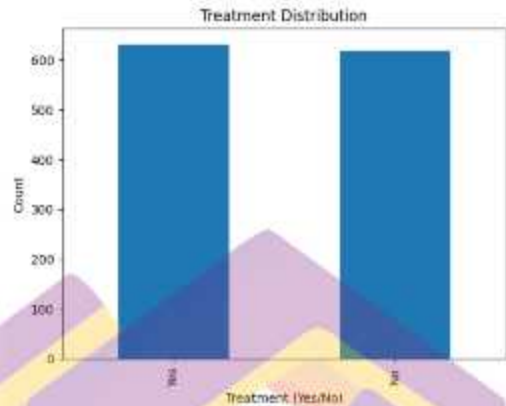


Diagram 4.4 Distribusi Memperoleh Perawatan

Sedangakan pada diagram di atas, pekerja yang mendapatkan perawatan dan tidak mendapatkan perawatan hampir seimbang. Hal ini menunjukan kalau hampir 50% perusahaan belum melakukan pengecekan terhadap kesehatan pekerja yang berkaitan dengan kesehatan mental, atau sebaliknya karyawan itu sendiri yang tidak memperdulikan kesehatan mentalnya sendiri.

4.2. Spliting Data

Rasio perbandingan pada spliting data mengacu pada pembagian dataset menjadi dua bagian dengan proporsi tertentu. Dalam hal ini 80% dari dataset digunakan untuk data training, sementara 20% sisanya digunakan untuk data testing. Rasio ini umumnya digunakan dan dianggap optimal untuk memastikan model memiliki akurasi yang baik tanpa mengorbankan ukuran dataset yang cukup besar untuk evaluasi. Dengan rasio 80:20 diharapkan dapat memastikan model terlatih dengan baik dan memiliki data yang cukup untuk menguji kinerjanya. Perbandingan yang tepat antara data latih dan data uji membantu memastikan bahwa tidak terjadi overfiting pada data latih. Perbandingan yang seimbang membantu menjaga integritas model dan memastikan bahwa model tersebut mencapai rasio terbaik dan siap untuk digunakan dalam kehidupan.

Akan tetapi demi memperoleh model yang mempunyai nilai akurasi yang optima, pada peneltian ini akan melakukan pengjian dengan rasio yang lain untuk memastikan manakah rasio yang memiliki performa terbaik dari model yang dibuat. Pada penelitian ini akan menggunakan rasio 90:10, 80:20, 70:30, 60:40 dan 50:50 dengan berberapa penyesuaian.

4.3. Machine Learning Scenario

4.3.1. Algoritma Random Forest

Berdasarkan spliting data di atas pada pengujian skenario pertama menggunakan algoritma random forest menggunakan dua hyperparameter tuning Grid Search dan Random Search yang masing-masing memiliki karakteristik tersendiri dalam perhitungannya. *Grid Search* mencoba semua kombinasi hyperparameter yang mungkin sedangkan *Random Search* memilih kombinasi hyperparameter secara acak yang biasanya efektif untuk hyperparameter dalam jumlah banyak (Pramudhyta and Rohman 2024b).

1. *Random Forest* dengan *Grid Search*

Pada pengujian ini *Random Forest* menggunakan *hyperparameter Grid Search* dengan dua parameter sebagai berikut :

Tabel 4.1. Tabel Parameter *Grid Search*

Parameter 1	Parameter 2
'n_estimators': [200, 300, 400],	'n_estimators': [100, 200, 300],
'max_depth': [10, 20, 30],	'max_depth': [10, 20, 30],
'min_samples_split': [2, 4, 6],	'min_samples_split': [2, 4, 6],
'min_samples_leaf': [1, 2, 4]	'min_samples_leaf': [1, 2, 4]

Perbedaan dari dua parameter 1 dan parameter 2 terdapat pada bagian estimator yang merupakan jumlah pohon keputusan untuk melihat parameter mana yang menghasilkan performa terbaik.

Gambar 4.3 di bawah ini adalah sintak pengujian algoritma *Random Forest* dengan *hyperparameter Grid Search* dengan parameter 1 yang

memiliki nilai akurasi paling optimal sebesar 82,63% dengan spliting data 60:40.

```
#Random Forest Grid Search
# Spliting data 60:40 Parameter 1
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=42)
```

```
# Hyperparameter Tuning Grid Search
param_grid = {
    'n_estimators': [200, 300, 400],
    'max_depth': [10, 20, 30],
    'min_samples_split': [2, 4, 6],
    'min_samples_leaf': [1, 2, 4]
    }
rf = RandomForestClassifier(random_state=42, class_weight='balanced')
grid_search = GridSearchCV(rf, param_grid, cv=5, scoring='roc_auc', n_jobs=-1)
grid_search.fit(X_train, y_train)

best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test)
print("Best Parameters:", grid_search.best_params_)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Classification Report:")
print(classification_report(y_test, y_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))
```

Gambar 4.3. *Grid Search* Spliting Data 60:40 Parameter 1

Adapun gambar 4.4 di bawah ini adalah sintak pengujian algoritma *Random Forest* dengan *hyperparameter Grid Search* dengan parameter 2 yang memiliki nilai akurasi paling optimal sebesar 82,63% dengan spliting data 60:40.

```
#Random Forest Grid Search
# Spliting data 60:40 Parameter 2
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=42)
```

```
# Hyperparameter Tuning Grid Search
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [10, 20, 30],
    'min_samples_split': [2, 4, 6],
    'min_samples_leaf': [1, 2, 4]
    }
rf = RandomForestClassifier(random_state=42, class_weight='balanced')
grid_search = GridSearchCV(rf, param_grid, cv=5, scoring='roc_auc', n_jobs=-1)
grid_search.fit(X_train, y_train)

best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_test)
print("Best Parameters:", grid_search.best_params_)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Classification Report:")
print(classification_report(y_test, y_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))
```

Gambar 4.4. *Grid Search* Spliting Data 60:40 Parameter 2

Dari beberapa sekema pengujian di atas, tabel 4.3 di bawah ini adalah hasil rekap nilai akurasi pada setiap pengujian algoritma *Random Forest* dengan *hyperparameter Grid Search*.

Tabel 4.2. Tabel Nilai Akurasi *Grid Search*

Dataset	Rasio	Akurasi Grid Search	
		Parameter 1	Parameter 2
Mental Health	90:10	78.57%	78.57%
	80:20	80.47%	80.47%
	70:30	81.91%	82.44%
	60:40	**82.63%**	**82.63%**
	50:50	82.26%	82.26%

2. *Random Forest* dengan *Random Search*

Pada pengujian ini, *Random Forest* menggunakan *hyperparameter Random Search* dengan 1 parameteryang disajikan dalam tabel 4.3 di bawah in. Akan tetapi karena pada pengujian pertama nilai akurasinya tidak kurang maksimal, maka pada penelitian ini menggunakan 2 pengujian dengan beberapa penyesuaian.

Tabel 4.3. Tabel Parameter *Random Search*

Parameter	Keterangan
'n_estimators': [1000, 200, 300, 4000, 500],	Jumlah pohon
'max_depth': [None, 10, 20, 30, 40],	Kedalaman maksimal
'min_samples_split': [2, 5, 10],	Minimal sampel untuk split node
'min_samples_leaf': [1, 2, 4],	Minimal sampel dalam leaf node
'bootstrap': [True, False]	Metode sampling

Gambar 4.5 di bawah ini adalah sintak yang menghasilkan nilai akurasi paling optimal pada pengujian pertama dengan nilai akuras 50% pada spliting data 80:20.

```
# Random Forest Random Search dengan Spliting data 90:10
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=42)

# Hyperparameter untuk RandomizedSearchCV
param_dist = {
    'n_estimators': [100, 200, 300, 400, 500],  # Jumlah pohon
    'max_depth': [None, 10, 20, 30, 40],  # Kedalaman maksimum
    'min_samples_split': [2, 5, 10],  # Minimal sampel untuk split node
    'min_samples_leaf': [1, 2, 4],  # Minimal sampel dalam leaf node
    'bootstrap': [True, False]  # Metode sampling
}

# Model RandomForest dengan class_weight='balanced' jika dataset tidak seimbang
rf = RandomForestClassifier(random_state=42, class_weight='balanced')

# RandomizedSearchCV dengan 20 iterasi dan validasi silang (cv=5)
random_search = RandomizedSearchCV(
    estimator=rf, param_distributions=param_dist,
    n_iter=20, cv=5, scoring='roc_auc', n_jobs=-1, random_state=42
```

Gambar 4.5. Random *Search* Pengujian ke-1

Rekap hasil pengujian *Random Forest* dengan *Random Search* pada pengujian pertama dapat dilihat pada tabel 4.4 di bawah ini :

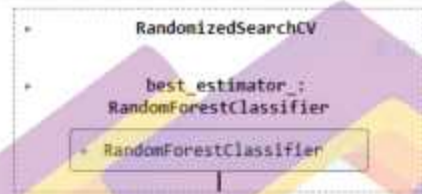
Tabel 4.4. Tabel Parameter *Grid Search*

Dataset	Rasio	Random Forest Random Search 1
Mental Health	90:10	49,2 %
	80:20	**50 %**
	70:30	44,7 %
	60:40	44,62 %
	50:50	44,76 %

Sedangkan gambar 4.6 adalah sintak yang menghasilkan nilai akurasi paling optimal pada pengujian ke-dua. Pada pengujian ini spliting data dilakukan internal oleh *cross validation* bukan dilakukan manual seperti pada sintak pertama. Untuk mendapatkan hasil yang optimal penulis melakukan kombinasi pada nilai k yaitu 5, 7, dan 9 serta pada kombinasi acaknya mulai dari 20, 50, dan 100 kombinasi. Dalam hal ini, nilai akurasi paling optimal diperoleh pada 100 kombinasi acak dan nilai k=5 dengan nilai akurasi sebesar 82.83%

```
# Randomized Search dengan 20 iterasi dan validasi silang (cv=5)
random_search = RandomizedSearchCV(
    estimator=rf, param_distributions=param_dist,
    n_iter=100, cv=5, scoring='roc_auc', n_jobs=-1, random_state=42
)
```

```
# Melatih model dengan RandomizedSearchCV
random_search.fit(X_train, y_train)
```




```
# Mengambil model terbaik
best_model = random_search.best_estimator_
```

Gambar 4.6. Random *Search* Pengujian ke-2

Setelah dilakukan perbaikan pada pengujian ke-dua seperti pada gamabr di atas, berikut penulis sajikan rekapan data nilai akurasi pada tabel 4.5 di bawah ini :

Tabel 4.5. Tabel Nilai Akurasi *Random Search* Pengujian 2

Dataset	Kombinasi	k=x	Akurasi
Mental Health	20	5	81,63
	20	7	82,44
	20	9	82,18
	50	5	82,43
	50	7	82,23
	50	9	82,44
	100	**5**	**82,83**
	100	7	82,44

	100	9	82,44

4.3.2. Algoritma AdaBoost

Setelah pada pengujian pertama menggunakan model *Random Forest* yang di optimasi dengan *hyperparameter Grid Search* dan *Random Search* dengan berbagai penyesuaian, pada pengujian yang ke-dua kali ini menggunakan algoritma *AdaBoost.* Untuk teknik pengujiannya sama dengan pengujian pertama yaitu menggunakan kombinasi spliting data 90:10, 80:20, 70:30, 60:40 dan 50:50. Pada gambar 4.7 di bawah ini menampilkan sintak yang memiliki nilai akurasi paling optimal sebesar 84,06% pada spiting data 80:20.

```
# Standardizing data
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

# AdaBoost Model Spliting Data 80:20
ada = AdaBoostClassifier(n_estimators=100, random_state=42)
ada.fit(X_train, y_train)
y_pred_ada = ada.predict(X_test)
print("AdaBoost Accuracy:", accuracy_score(y_test, y_pred_ada))
print(classification_report(y_test, y_pred_ada))

AdaBoost Accuracy: 0.8406374501992032
              precision    recall  f1-score   support

           0       0.88      0.79      0.83       124
           1       0.81      0.90      0.85       127

    accuracy                           0.84       251
   macro avg       0.85      0.84      0.84       251
weighted avg       0.84      0.84      0.84       251
```

Gambar 4.7. Sintak *AdaBoost* Spliting Data 80:20

Tabel 4.6 di bawah ini adalah rekapan hasil nilai akurasi *AdaBoost* pada setiap kombinasi spliting data.

Tabel 4.6. Tabel Nilai Akurasi *AdaBoost*

Dataset	Rasio	AdaBoost
Mental Health	90:10	78,89 %
	80:20	**84,06 %**
	70:30	81,64 %
	60:40	82,35 %

	50:50	81,23 %

4.3.3. Algoritma XGBoost

Pengujian yang ke-tiga adalah menggunakan algoritma *XGBoost*. Pada pengujian ini, kombinasi yang digunakan sama dengan pengujian sebelumnya yaitu dengan kombinasi spliting data 90:10, 80:20, 70:30, 60:40 dan 50:50. Sintak pada gambar 4.8 di bawah ini adalah kombinasi spliting data 80:20 yang memiliki nilai akurasi paling optimal sebesar 80,47%.

Gambar 4.8. Sintak *XGBoost* Spliting Data 80:20

Rekapan data nilai akuasi algoritma XGBoost pada masing-masing splting data, dirangkum pada tabel 4.7 di bawah ini :

Tabel 4.7. Tabel Nilai Akurasi *XGBoost*

Dataset	Rasio	XGBoost
	90:10	79,08 %
	80:20	**80,47 %**
Mental Health	70:30	77,65 %
	60:40	80,43 %
	50:50	80,17 %

4.3.4. Kombiasi Algoritma *Random Forest* dan *AdaBoost*

Pada pengujian ke-4 ini, penulis mencoba mengkombinasikan algoritma *Random Forest* dan *AdaBoost*, harapannya dengan kombinasi ini diperoleh nilai yang lebih optimal dibandingkan model yang dibuat sebelumnya. Akan tetapi hasil yang diperoleh masih belum begitu optimal. Spliting data pada pengujian kali ini masih sama seperti pengujian sebelumnya yaitu kombinasi 90:10, 80:20, 70:30, 60:40 dan 50:50. Pada gambar 4.9 di bawah ini, ditampilkan sintak kombinasi *Random Forest* dan *AdaBoost* yang mempunyai niai akurasi paling optimal sebesar 83,66% pada spliting data 80:20

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42

# Random Forest + AdaBoost
rf_ada = RandomForestClassifier(n_estimators=500, random_state=42)
rp_ada_boost = AdaBoostClassifier(estimator=rf_ada, n_estimators=500, random_state=42)
rp_ada_boost.fit(X_train, y_train)
y_pred_ada = rp_ada_boost.predict(X_test)
print("Random Forest + AdaBoost Accuracy:", accuracy_score(y_test, y_pred_ada))
print(classification_report(y_test, y_pred_ada))

Random Forest + AdaBoost Accuracy: 0.8366533864541833
              precision    recall  f1-score   support

           0       0.85      0.81      0.83       124
           1       0.82      0.87      0.84       127

    accuracy                           0.84       251
   macro avg       0.84      0.84      0.84       251
weighted avg       0.84      0.84      0.84       251
```

Gambar 4.9. Sintak *Random Forest* dan *AdaBoost* Spliting Data 80:20

Lebih detail rekapan nilai hasil akurasi kombinasi *Random Forest* dan *AdaBoost* pada setiap spliting data disajikan pada tabel 4.8 di bawah ini.

Tabel 4.8. Tabel Nilai Akurasi *Random Forest dan AdaBoost*

Dataset	Rasio	Random Forest AdaBoost
Mental Health	90:10	78,77 %
	80:20	**83,66 %**
	70:30	82,63 %
	60:40	80,48 %
	50:50	82,67 %

4.3.5. Kombinasi Algoritma *Random Forest* dan *XGBoost*

Setelah pada pengujian ke-4 melakukan kombinasi pengujian Random Forest dan AdaBoost belum memperoleh nilai akurasi yang signifikan, pada pengujian ke-5 ini penguji melakukan pengujian pada kombinasi Random Forest dan XGBoost. Akan tetapi hasil pengujian ke-5 ini tidak lebih baik dari pada pengujian ke-4. Perlakuannya masih sama seperti pada pengujian sebelumya yaitu melakukan kombinasi splitng data 90:10, 80:20, 70:30, 60:40 dan 50:50. Sintak yang di tampilkan pada gambar 4.9 di bawah ini adalah yang memiliki nilai akurasi paling optimal sebesar 80,475 pada pembagian spliting data 80:20.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

# Random Forest + XGBoost
rf_xgb = RandomForestClassifier(n_estimators=100, random_state=42)
rf_xgb.fit(X_train, y_train)
xgb = XGBClassifier(n_estimators=50, random_state=42, use_label_encoder=False, eval_metric='logloss')
xgb.fit(X_train, y_train)
y_pred_xgb = xgb.predict(X_test)
print("Random Forest + XGBoost Accuracy:", accuracy_score(y_test, y_pred_xgb))
print(classification_report(y_test, y_pred_xgb))
```

```
Random Forest + XGBoost Accuracy: 0.8047808764940239
              precision    recall  f1-score   support

           0       0.82      0.77      0.80       124
           1       0.79      0.83      0.81       127

    accuracy                           0.80       251
   macro avg       0.81      0.80      0.80       251
weighted avg       0.81      0.80      0.80       251
```

Gambar 4.10. Sintak *Random Forest* dan *AdaBoost* Spliting Data 80:20

Rekapan lengkap nilai akurasi pada pengujian ke-5 yang melakukan kombinasi algoritm *Random Forest* dan *XGBoost* tersaji dalam tabel 4.9 di bawah ini :

Tabel 4.9. Tabel Nilai Akurasi *Random Forest dan XGBoost*

Dataset	Rasio	Random Forest XGBoost
Mental Health	90:10	79,78 %
	80:20	**80,47 %**
	70:30	80,23 %
	60:40	80,31 %
	50:50	78,88 %

4.3.6. Matriks Evaluasi Akurasi, *Precision, Recall,* dan *F1-Score*

Setelah pada pembahasan sebelumnya disajikan data rekap nilai akurasi pada semua pengujian, pada sub bab ini penulis akan rangkum nilai akurasi, *precision, recall,* dan *f1-score* yang paling optimal pada setiap pengujian. Rangkumannya tersaji pada tabel 4.10 di bawah ini.

Tabel 4.10. Tabel Nilai *Precison, Recall* dan *F1-Score*

No	Algoritma	Akurasi (%)	*Precision*(%)		*Recall* (%)		*F1-Score* (%)	
			0	1	0	1	0	1
1	Random Forest Grid Search	82,63	87	79	71	93	79	85
2	Random Forest Random Search 1	50	49	51	51	50	49	48
3	Random Forest Random Search 1	82,83	91	78	73	92	81	84
4	**AdaBoost**	**84,06**	**88**	**81**	**78**	**90**	**83**	**85**
5	XGBoost	80,47	84	78	75	86	79	82
6	Random Forest dan AdaBoost	83,66	85	82	81	87	83	84
7	Random Forest dan XGBoost	80,47	82	79	77	83	80	81

Tabel 4.10 yang menyajikan rekapan nilai *precision, recall,* dan *f1-score* sengaja ditampilkan dalam kelas 0 dan kelas 1. Tujuannya adalah untuk melihat seberapa optimal model yang dibuat pada masing-masing kelas. Setelah di ambil rata-rata, algoritma *AdaBoost* yang memiliki performas paling optimal sesuai dengan nilai akurasinya. Lebih detail pembahasan mengenai nilai akurasi, *precision, recall* dan *f1-score* ditampilkan pada tabel 4.11 di bawah ini :

Tabel 4.11. Tabel *Classification Report* Parameter 1

Classification Report:

Class	precision	recal	f1-score	Support
0	0.88	0.78	0.83	124
1	0.81	0.90	0.85	127
Accuracy			0.84	251

macro avg	0.85	0.84	0.84	251
Weighted avg	0.84	0.84	0.84	251

Keterangan :

Kelas 0 : Tidak menjalani pengobatan

Kelas 1 : Menjalani pengobatan

Penjelasan Tabel :

1. *Precision* (Presisi):
   - Untuk kelas 0 : dari semua prediksi kelas 0, 88% adalah benar.
   - Untuk kelas 1 : dari semua prediksi kelas 1, 81% adalah benar.
2. *Recall* (Sensitivitas):
   - Untuk kelas 0 : hanya 78% dari semua kelas 0 yang berhasil ditemukan model.
   - Untuk kelas 1 : 90% dari semua kelas 1 berhasil ditemukan model.
3. F1-*Score*: kombinasi antara precision dan recall.
   - Kelas 1 memiliki nilai *F1-Score* lebih tinggi sebesar 85% dibandingkan kelas 0 yang hanya 83%, yang artinya model lebih baik dalam mendeteksi kelas 1.
4. *Accuracy*
   - Akurasi dari model yang telah 84%. Artinya: dari 251 data, 84% diprediksi dengan benar.
5. *Macro avg dan Weighted avg*
   - Macro avg: rata-rata dari semua kelas tanpa memperhatikan proporsi (sama rata).
   - Weighted avg: rata-rata yang mempertimbangkan jumlah data tiap kelas (support).
6. *Confusion Matrix*

Interpretasi :

- 124 data kelas 0 diprediksi benar.
- 52 data kelas 0 salah dikira kelas 1 (*False Positive*).
- 184 data kelas 1 diprediksi benar.
- 16 data kelas 1 salah dikira kelas 0 (*False Negative*).

Kesimpulan

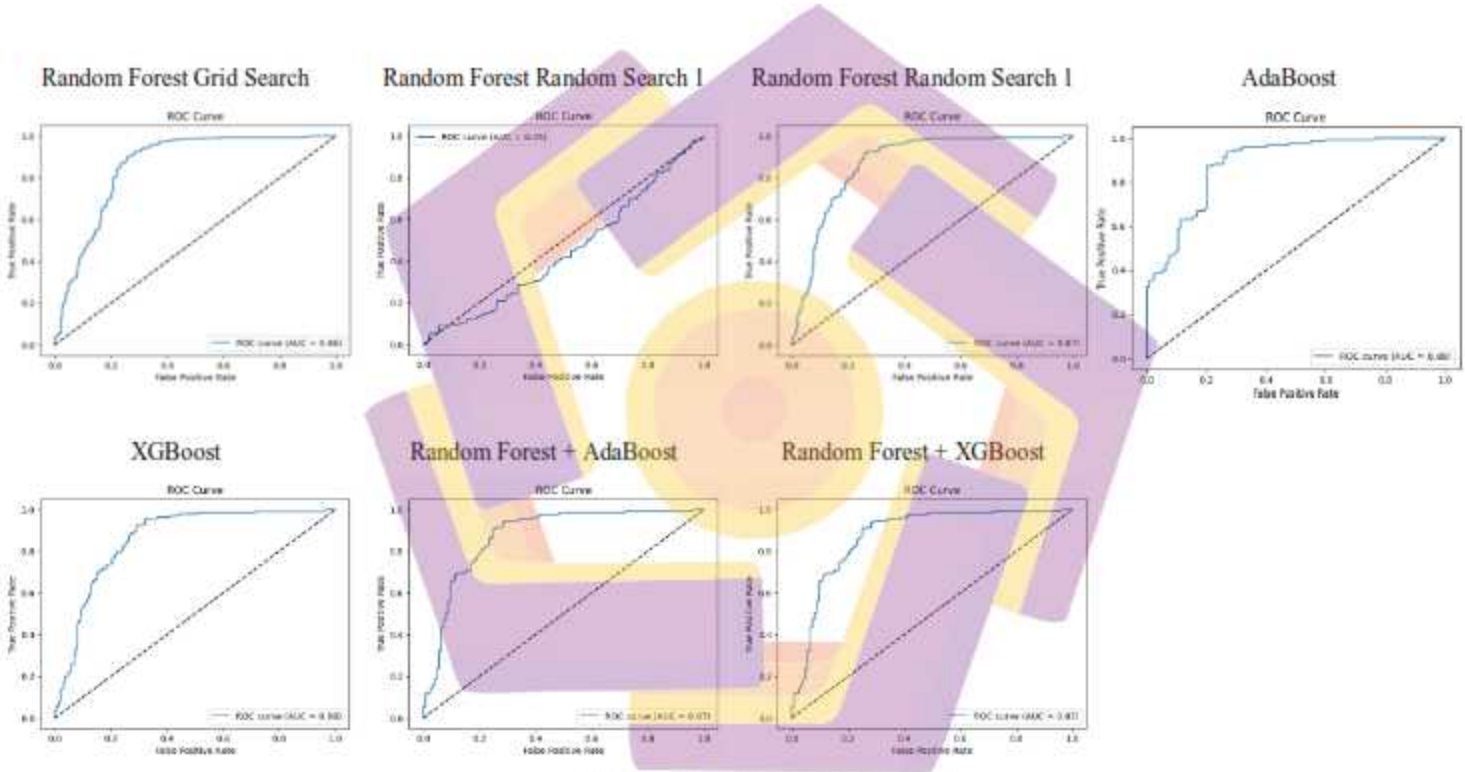
- Model secara umum cukup bagus dengan nilai akurasi sebesar 84%, dengan nilai *F1-Score* 85% pada kelas 1.
- Model masih banyak salah dalam mengenali kelas 0 meskipun tidak signifikan karena nilai F1-Score sebesar 83%.
- Recall untuk kelas 1 mencapai 90%, yang berarti hanya 10% data kelas 1 tidak terdeteksi.

4.3.7. Kurva ROC (*Receiver Operating Characteristic*)

Kurva ROC (*Receiver Operaing Characteristic*) merupakan grafik yang digunakan untuk mengevaluasi kinerja model klasifikasi, khususnya pada masalah klasifikasi biner (Pendrill et al. 2023). Kurva ini menggambarkan hubungan antara True Positive Rate (TPR) atau *Sensitivity* pada sumbu Y dan False Positive Rate (FPR) pada sumbu X, yang dihitung pada berbagai nilai ambang (*threshold*).

Kurva ROC menunjukkan kemampuan model dalam membedakan antara kelas positif dan kelas negatif. Semakin dekat kurva ROC ke sudut kiri atas grafik, semakin baik kinerja model klasifikasi tersebut.

Untuk mengkuantifikasi performa model dari kurva ROC, digunakan nilai AUC (*Area Under the Curve*). Nilai AUC berada pada rentang 0–1, di mana ketika AUC mendekati atau sama dengan 1 berarti model sangat baik (perfect classifier) sebaliknya jika nilai AUC mendekati nilai 0, maka model yang dibuat tidak optimal. Dalam sub bab ini akan menampilkan 7 kurva ROC pada masing-masing pengujian yang niai akurasinya paling optimal. Tujuannya adalah sebagai gambaran seberapa kinerja model yang telah dibuat dalam penelitian ini.

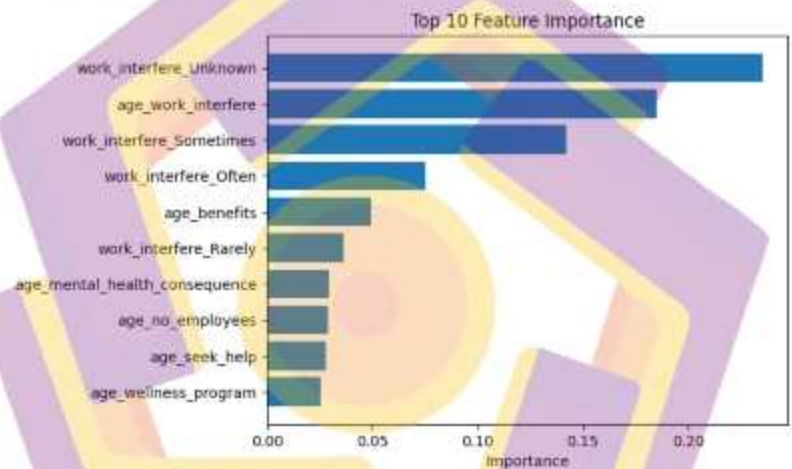


Gambar 4.11. Kurva ROC Setiap Pengujian

Dari kurva ROC di atasa, dapat di simpukan bahwa kinerja model yang dibuat paling optimal pada algoritma AdaBoost dengan nilai AUS sebesar 88%.

4.3.8. Sepuluh Kreiteria yang Paling Berpengaruh Terhadap Model

Setelah diperoleh kinerja model paling berdasarkan kura ROC adalah pada algoritma AdaBoost, maka perlu dilakukan pengecekan 10 kriteria yang paling berpegaruh terhadap kinerja model.



1. Faktor yang paling berpengaruh adalah **gangguan pekerjaan akibat kesehatan mental**, terutama jika informasinya tidak tersedia (**Unknown**).
2. **Usia** memainkan peran penting dalam berbagai faktor, seperti dampaknya terhadap pekerjaan, akses ke tunjangan, dan keputusan untuk mencari bantuan.
3. **Tunjangan kesehatan (benefits)** juga berpengaruh terhadap prediksi model.
4. Secara keseluruhan, model menemukan pola bahwa **gangguan pekerjaan akibat kesehatan mental dan akses terhadap tunjangan** memiliki hubungan kuat dengan hasil prediksi.

4.4. Tabel Evaluasi Menggunakan *K-Fold Cross Validation*

Berdasarkan masukan dari penguji untuk melakukan evaluasi model yang telah dibuat dalam penelitian ini menggunakan *k-fold cross-validation*, berikut adalah hasil yang diperoleh dengan nilai *k=5* dan dirangkum dalam tabel di bawah ini :

Tabel 4.12. Tabel Evaluasi dengan *K-Fold Cross Validation*

No	Algoritma	k=x	Akurasi
1	*Random Forest*	5	81,69
2	*AdaBoost*	5	82,65
3	*XGBoost*	5	79,37

Nilai akurasi paling optimal dengan evaluasi *k-fold cross-validation* sebesar 82,65 pada algoritma random forest. Maka dari itu , pada penelitian ini baik dengan spliting data maupun evauasi *k-fold cross-valitadion*, algoritma *AdaBoost* mengungguli algoritma yang lain.

4.5. Tabel Perbandingan Nilai Akurasi Dengan Penelitian Sebelumnya

Untuk mengetahui perbandingan penelitian ini dengan penelitian sebelumnya, berikut adalah tabel nilai akurasi yang paling optimal pada setiap penelitian yang telah diambil sebagai referensi pada penelitian ini. Berdasarakan tabel 4.13 di bawah ini, dapat diketahui bahwa nilai akurasi yang diperoleh pada penelitian ini paling optimal 84,06% pada algoritma *AdaBoost*. Bila dibandingkan dengan penelitian sebelumnya hanya mampu mengungguli satu penelitian yang dijadikan referensi pokok dalam penelitian ini. Berbeda untuk 5 penelitian yang lain, nilai akurasi lebih tinggi bila dibandingkan dengan penelitian ini.

Sedangakan untuk penelitian yang menggunakan dataset sama, nilai akurasi yang diperoleh meningkat 0,19% untuk algoritma *AdaBoost*. Sedangkan untuk algoritma *XGBoost* yang memiliki nilai paling optimal dalam penelitian sebelumnya sebesar 84,94%, nilai akurasi pada penelitian ini lebih rendah 0,88 dengan algoritma *AdaBoost*. Meskipun demikian, penelitian ini unggul dalam mendeteksi semua data pekerja yang membutuhkan perawatan (kelas 1) yang

terbukti dari niai recall yang mencapai 90%. Selain unggul dalam mendeteksi semua data terutama dalam kelas 1, pada penelitian ini juga disajikan faktor apa saja yang yang mempengaruhi perhitungan model yang dibuat. Secara tidak langsung, ini 10 faktor ini bisa dijadikan perbaikan dalam perusahaan agar karyawan tidak mengalami gangguan kesehatan mental yang diakibatkan dari 10 faktor tersebut. Sehingga penanganan gangguan kesehatan mental bisa dilakukan dari dua sisi, yang pertama penanganan dari segi medis dan penanganan kedua dari kebijakan perusahaan berdasarkan faktor-faktor yang disebutkan dalam penelitan ini.

Secara umum, referensi pokok yang penulis gunakan pada tabel 4.13 di bawah relatif digunakan karena memiliki kesamaan pada kasus yang diteliti yaitu mengenai kesehatan mental. Referensi yang yang memiliki kesamaan hanya terdapat pada poin nomor 2 yang menggunakan dataset sama dan algoritma yang kurang lebih sama. Sehingga

Tabel 4.13. Tabel Referensi Pokok Pada Kasus Kesehatan Mental

No	Judul	Nama Peneliti dan Tahun	Algoritma	Akurasi
1	Detection of depression and anxiety in the perinatal period using Marine Predators Algorithm and kNN	Nur Banu Oğur, Muhammed Kotan, Deniz Balta, Burcu Çarklı Yavuz, Yavuz Selim Oğur, Hilal Uslu Yuvacı, Esra, Yazıcı (2023)	MPA+kNN	98,11%
2	Mental Health Analysis of Employees using Machine Learning Techniques	Sujal BH, Neelima K, Deepanjali C, Bhuvanashree P, Kavitha Duraipandian, Sharanya Rajan, Mithileysh Sathiyanarayanan, (2022)	XGBoost	84,94%
3	Prediction of Mental Health Among University Students	Fahruddin Sahlan, Faris Hamidi, M. Zulhafizal Misrat, M. Haziq Adli, Sharyar Wani, Yonis Gulzar (2021)	Decision Trees	64%
4	Comparison Of Random Forest, Support Vector Machine And Naive Bayes Algorithms To Analyze Sentiment Towards Mental Health Stigma	Putri Elisa, Auliya Rahman Isnain (2024)	SVM	86,11%
5	Klasifikasi Mental Mahasiswa Menggunakan Metode Machine Learning	Nirwan Moningka, Raynold, M. Hafidurrohman, Wahyu Ajri Tri R, Kusrini (2023)	KNN	85%
6	Classification of Employee Mental Health Disorder Treatment with K-Nearest Neighbor Algorithm	Hakkun Elmunsyah, Risalatul Mu'awanah, Triyanna Widiyaningtyas (2019)	KNN	87,27
7	Optimasi Algoritma Random Forest Untuk Diagnosis Gangguan Kesehatan Mental	Riswanto (2025)	AdaBoost	84,06%

BAB V

PENUTUP

5.1. Kesimpulan

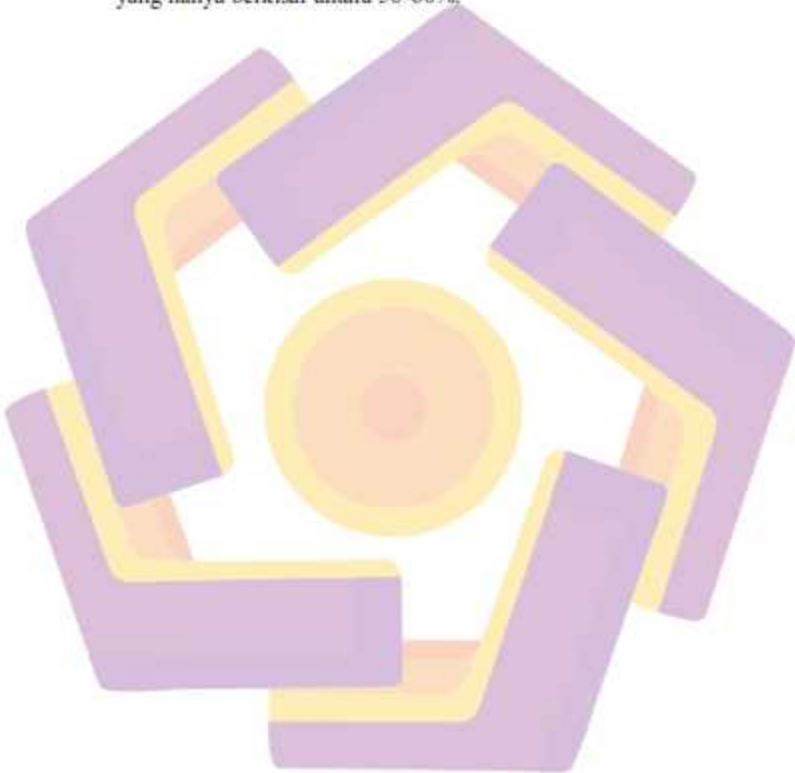
Berdasarakan hasil dari pembahasan pada bab sebelumnya diperoleh nilai akurasi tertinggi menggunakan model *AdaBoost* dengan nilai akurasi sebesar 84.06%, diikuti dengan *Random Forest Grid Search* dengan 82.63%, *Random Forest + AdaBoost* sebesar 82.47%, *XGBoost* dan *Random Forest + XGBoost* sebesar 80.47% dan terakhir *Random Forest Random Search* dengan nilai akurasi antara 50-60%. Selanjutnya peneliti melakukan pengujian pada *Random Search* dengan sedikit perubahan yaitu menghilangkan splitng data yang artinya dataset 100% digunakan sebagai data training dan data testing serta merubah nilai k dan jumah kombinasinya. Dengan perubahan tersebut diperoleh nilai akurasi paling optimal sebesar 82,83% pada 100 kombinasi dan nilai k=5. Hasil yang diperoleh menggunakan berbagai hyperparameter yang disesuaikan dengan karakteristrik dari setiap model sehingga diperoleh hasil paling optimal pada setiap modelnya.

5.2. Saran

Berdasarkan hasil yang diperoleh, masih perlu banyak perbaikan terutama untuk meningkatkan akurasi pada setiap model karena dalam penelitian ini hanya fokus membahas optimasi random forest. Berikut adalah beberapa saran yang nantinya bisa digunakan untuk meningkatkan hasil perhitungan random forest :

3. Penggunaan hyperparameter lain sangat dibutuhkan agar perhitungan random forest bisa lebih optimal.
4. Coba menggunakan boosting yang lebih canggih agar nilai akurasinya juga meningkat
5. Lakukan pembersihan fitur pada dataset sehingga meminimalkan terjadinya noice.

6. Jika akan menganalisis dataset yang sama akan lebih baik jika dilakukan perbandingan dengan algoritma yang lain sehingga bisa diperbandingkan algoritma mana yang memiliki nilai akurasi paling optimal.
7. Perlu dilakukan kajian lebih mendalam mengenai nilai akurasi dengan menggunakan Random Search dengan spliting data karena nilai akurasi yang hanya berkisar antara 50-60%.



DAFTAR PUSTAKA

PUSTAKA BUKU

Saputra, Irwansyah. (2023). Belajar Mudah Data Mining untuk Pemula. INFORMATIKA. Bandung

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Alfaresy Chaerudin, Reza, Ruth Mariana Bunga Wadu, Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas RS Pembangunan Nasional Veteran Jakarta Jl Fatmawati Raya, Pd Labu, Kec Cilandak, Kota Depok, and Jawa Barat. 2022. *Implementasi Algoritma Naïve Bayes Untuk Analisis Klasifikasi Survei Kesehatan Mental (Studi Kasus: Open Sourcing Mental Illness)*.

Bh, Sujal, K. Neelima, C. Deepanjali, P. Bhuvanashree, Kavitha Duraipandian, Sharanya Rajan, and Mithileysh Sathiyanarayanan. 2022. "Mental Health Analysis of Employees Using Machine Learning Techniques." Pp. 1–6 in *2022 14th International Conference on COMmunication Systems and NETworkS, COMSNETS 2022*. Institute of Electrical and Electronics Engineers Inc.

Britanthia, Lukhia, Christina Tanujaya, Bambang Susanto, and Asido Saragih. 2020. "Perbandingan Metode Regresi Logistik Dan Random Forest Untuk Klasifikasi Fitur Mode Audio Spotify." *Indonesian Journal of Data and Science (IJODAS)* 1(3):68–78.

Elisa, Putri, and Auliya Rahman Isnain. 2024. "COMPARISON OF RANDOM FOREST, SUPPORT VECTOR MACHINE AND NAIVE BAYES

ALGORITHMS TO ANALYZE SENTIMENT TOWARDS MENTAL HEALTH STIGMA." *Jurnal Teknik Informatika (JUTIF)* 5(1):321–29. doi:10.52436/1.jutif.2024.5.1.1817.

Elmunsyah, Hakkun, Risalatul Mu'awanah, Triyanna Widiyaningtyas, Ilham A. E. Zaeni, and Felix Andika Dwiyanto. 2019. "Classification of Employee Mental Health Disorder Treatment with K-Nearest Neighbor Algorithm." Pp. 211–15 in *ICEEIE 2019 - International Conference on Electrical, Electronics and Information Engineering: Emerging Innovative Technology for Sustainable Future*. Institute of Electrical and Electronics Engineers Inc.

Erkamim, Moh., Adam M. Tanniewa, Irfan AP, and Nurhayati Nurhayati. 2024. "Analisis Perbandingan Metode AdaBoost, Gradient Boosting, Dan XGBoost Untuk Kalsifikasi Status Gizi Pada Balita." *Building of Informatics, Technology and Science (BITS)* 6(3):1799–1807. doi:10.47065/bits.v6i3.5717.

Febianto, Resta Trias, Dewi Suranti, and Rizka Tri Alinse. 2024. *PENERAPAN ALGORITMA ADABOOST DALAM MENGETAHUI POLA PENGGUNA KB DI PUSKESMAS TANJUNG HARAPAN*. http://jurnal.goretanpena.com/index.php/JSSR.

Geetha, Dinesh, Indhusree S, Kavitha S N, and Ankush K Gupta. 2023. "Development of Mental HealthWebsite and Securing the Chat Application Using 4X4 Hill Cipher Technique." *Journal of Harbin Engineering University* 44(7).

Herrman, Helen., Shekhar. Saxena, Rob. Moodie, World Health Organization. Department of Mental Health and Substance Abuse., Victorian Health Promotion Foundation., and University of Melbourne. 2005. *Promoting Mental Health : Concepts, Emerging Evidence, Practice*. World Health Organization.

Idaiani, Sri, and Edduwar Idul Riyadi. 2018. "Sistem Kesehatan Jiwa Di Indonesia: Tantangan Untuk Memenuhi Kebutuhan." *Jurnal Penelitian Dan Pengembangan Pelayanan Kesehatan* 70–80. doi:10.22435/jpppk.v2i2.134.

Intan Permata, and Esther Sorta Mauli Nababan. 2023. "Analisis Perbandingan Algoritma XGBoost Dan Algoritma Random Forest Ensemble Learning Pada Klasifikasi Keputusan Kredit." *JURNAL RISET RUMPUN MATEMATIKA DAN ILMU PENGETAHUAN ALAM* 2(2):65–71. doi:10.55606/jurrimipa.v2i2.1336.

Kang, Mingu, Siho Shin, Jaehyo Jung, and Youn Tae Kim. 2021. "Classification of Mental Stress Using CNN-LSTM Algorithms with Electrocardiogram Signals." *Journal of Healthcare Engineering* 2021. doi:10.1155/2021/9951905.

Li, Jessie. 2023. *Asymptotics of K-Fold Cross Validation*. Vol. 78.

Mahlich, Christopher, Tobias Vente, and Joeran Beel. 2024. *From Theory to Practice: Implementing and Evaluating e-Fold Cross-Validation*.

Maulida, Vina, Rudy Herteno, Dwi Kartini, Friska Abadi, and Mohammad Reza Faisal. 2023. "Feature Selection Using Firefly Algorithm With Tree-Based Classification In Software Defect Prediction." *Journal of Electronics,*

Electromedical Engineering, and Medical Informatics 5(4). doi:10.35882/jeeemi.v5i4.315.

Moningka, Nirwan, M. Hafidurrohman, and Wahyu R. Ajri Tri. 2023. *Klasifikasi Mental Mahasiswa Menggunakan Metode Machine Learning*. Vol. 1. https://www.kaggle.com/datasets/shariful07/student-mental-.

Muslim Karo Karo, Ichwanul. 2020. *Implementasi Metode XGBoost Dan Feature Importance Untuk Klasifikasi Pada Kebakaran Hutan Dan Lahan*. Vol. 1.

Negeri, Universitas, Surabaya Jl, Lidah Lidah Wetan, Kec Wetan, Kota Lakarsantri, Jawa Surabaya, and Indonesia Timur. 2023. *Komparasi Algoritma Random Forest, Naïve Bayes, Dan Bert Untuk Multi-Class Classification Pada Artikel Cable News Network (CNN) Nanang Huxin*. Vol. 7.

Nguyen, Quoc Huy, Anh Tuan Tran, Thi Ngoc-Thanh Nguyen, and Nhu Tai Do. 2024. "A Meta-Heuristic Approach for Enhancing Performance of Associative Classification." *ICT Research* 47–53. doi:10.32913/mic-ict-research.v2024.n1.1246.

Nisa, Khoirun. 2024. "Klasifikasi Penyakit Gangguan Mental Dengan Algoritma LightGBM." *Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)* 9(2):1086–94. https://tunasbangsa.ac.id/ejurnal/index.php/jurasik.

Novianti, Nia, Muhammad Zarlis, and Poltak Sihombing. 2022. "Penerapan Algoritma Adaboost Untuk Peningkatan Kinerja Klasifikasi Data Mining

Pada Imbalance Dataset Diabetes." *JURNAL MEDIA INFORMATIKA BUDIDARMA* 6(2):1200. doi:10.30865/mib.v6i2.4017.

Ogunseye, Elizabeth Oluyemisi, Cecilia Ajowho Adenusi, Andrew C. Nwanakwaugwu, Sunday Adeola Ajagbe, and Solomon O. Akinola. 2022. "Predictive Analysis of Mental Health Conditions Using AdaBoost Algorithm." *ParadigmPlus* 3(2):11–26. doi:10.55969/paradigmplus.v3n2a2.

Oğur, Nur Banu, Muhammed Kotan, Deniz Balta, Burcu Çarklı Yavuz, Yavuz Selim Oğur, Hilal Uslu Yuvacı, and Esra Yazıcı. 2023. "Detection of Depression and Anxiety in the Perinatal Period Using Marine Predators Algorithm and KNN." *Computers in Biology and Medicine* 161. doi:10.1016/j.compbiomed.2023.107003.

Oktafiani, Rian, Arief Hermawan, and Donny Avianto. 2023. "Pengaruh Komposisi Split Data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning." *Jurnal Sains Dan Informatika* 19–28. doi:10.34128/jsi.v9i1.622.

Parinduri, Syawaluddin Kadafi, Putrama Alkhairi, and Hendry Qurniawan. 2025. "Classification Model Optimization Using Grid Search and Random Search in Machine Learning Algorithms." *Bulletin of Informatics and Data Science* 4(2):71–78. doi:10.61944/bids.v4i2.136.

Pramudhyta, Nirvan Adam, and Muhammad Syaifur Rohman. 2024a. "Perbandingan Optimasi Metode Grid Search Dan Random Search Dalam Algoritma XGBoost Untuk Klasifikasi Stunting." *JURNAL MEDIA INFORMATIKA BUDIDARMA* 8(1):19. doi:10.30865/mib.v8i1.6965.

Pramudhyta, Nirvan Adam, and Muhammad Syaifur Rohman. 2024b. "Perbandingan Optimasi Metode Grid Search Dan Random Search Dalam Algoritma XGBoost Untuk Klasifikasi Stunting." *JURNAL MEDIA INFORMATIKA BUDIDARMA* 8(1):19. doi:10.30865/mib.v8i1.6965.

Primadhani Tirtopangarsa, Arianti, and Warih Maharani. 2021. "Sentiment Analysis of Depression Detection on Twitter Social Media Users Using the K-Nearest Neighbor Method Analisis Sentimen Deteksi Depresi Pada Pengguna Media Sosial Twitter Dengan Menggunakan Metode K-Nearest Neighbor." 13–2021.

Rahmayani, Ririt Sheila Tina, and Fikri Budiman. 2024. "Analisa Optimasi Grid Search Pada Algoritma Random Forest Dan Decision Tree Untuk Klasifikasi Stunting." *Building of Informatics, Technology and Science (BITS)* 6(3). doi:10.47065/bits.v6i3.6128.

Rezk, Sahar Saeed, and Kamal Samy Selim. 2024. "Metaheuristic-Based Ensemble Learning: An Extensive Review of Methods and Applications." *Neural Computing and Applications* 36(29):17931–59.

Rudi Fanani, M., and Elvinda Bendra Agustina. 2024. *Journal of Artificial Intelligence and Engineering Applications Implementation Of the C4.5 Algorithm With The Backward Elimination Feature Selection For MSME Product Sales Strategy*. Vol. 3. https://ioinformatic.org/.

Sahlan, Fadhluddin, Faris Hamidi, Muhammad Zulhafizal Misrat, Muhammad Haziq Adli, Sharyar Wani, and Yonis Gulzar. 2021. *Prediction of Mental Health Among University Students*. Vol. 7.

Salman, Hasan Ahmed, Ali Kalakech, and Amani Steiti. 2024. "Random Forest Algorithm Overview." *Babylonian Journal of Machine Learning* 2024:69–79.

Saragih, Triando Hamonangan, M. Reza Faisal, Dan Muhammad Al, Ichsan Nur, and Rizqi Said. 2021. *AdaBoost Classifier Untuk Klasifikasi Tanaman Jarak Pagar*. Vol. 9.

Siraj-Ud-Doulah, Md, and Md Ashad Alam. 2018. *Ecological Data Analysis Based on Machine Learning Algorithms*.

Tentua, Meilany Nonsi, Vicky Fidiantoro, and Pradana Feri Ariyanto. 2022. "Metode C4 Metode C4.5 .5 Dan Naive." *Jurnal Dinamika Informatika* 11(2). https://www.kaggle.com/osmi/mental-.

Vaishnavi, Konda, U. Nikhitha Kamath, B. Ashwath Rao, and N. V. Subba Reddy. 2022. "Predicting Mental Health Illness Using Machine Learning Algorithms." in *Journal of Physics: Conference Series*. Vol. 2161. IOP Publishing Ltd.

White, Jacob, and Sarah D. Power. 2023. "K-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation." *Sensors* 23(13). doi:10.3390/s23136077.

World Health Organization. 2022. "World Mental Health Report :Transforming Mental Health for All." https://www.who.int/teams/mental-health-and-substance-use/world-mental-health-report.

Yaqoob, Abrar, Navneet Kumar Verma, Mushtaq Ahmad Mir, Ghanshyam G. Tejani, Nashwa Hassan Babiker Eisa, Hind Mamoun Hussien Osman, and Mohd Asif Shah. 2025. "SGA-Driven Feature Selection and Random Forest Classification for Enhanced Breast Cancer Diagnosis: A Comparative Study." *Scientific Reports* 15(1). doi:10.1038/s41598-025-95786-1.

Yulianti, Elina Herni, Oni Soesanto, and Yuana Sukmawaty. 2022. "Penerapan Metode Extreme Gradient Boosting (XGBOOST) Pada Klasifikasi Nasabah Kartu Kredit." *JOMTA Journal of Mathematics: Theory and Applications* 4(1).

Yustikasari, Renata Anisa, and Retasari Dewi. 2022. "Pemanfaatan Program Implementasi Promosi Kesehatan: Promosi Kesehatan Mental Pada Remaja." *Jurnal Pengabdian Masyarakat* Vol. 1(No. 3).

Zeng, Guoping. 2020. "On the Confusion Matrix in Credit Scoring and Its Analytical Properties." *Communications in Statistics - Theory and Methods* 49(9):2080–93. doi:10.1080/03610926.2019.1568485.

PUSTAKA LAPORAN PENELITIAN

Nama peneliti, tahun, judul, jenis penelitian, nama lembaga, kota

Siti, Kalimah, 2022, Klasifikasi Penyakit Diabetes Menggunakan Metode Decision Tree dan Random Forest, Skripsi, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sriwijaya, Palembang

Kurniawan, M. P., 2011, Teknologi Motion Capture dengan Multi Kamera pada Pembuatan Animasi 3D, Tesis, Magister Teknik Informatika, STMIK AMIKOM, Yogyakarta

PUSTAKA ELEKTRONIK

Nama penulis, tanggal akses, judul artikel, alamat URL secara lengkap. Publikasi di web selain e-book, e-journal, dan e-proceeding tidak diperbolehkan untuk dijadikan rujukan penelitian ilmiah

Utami, E.; Istiyanto, J.E.; Hartati, S.; Marsono; Ashari, A., 25 November 2009, Developing Transliteration Pattern of Latin Character Text Document Algorithm Based on Linguistics Knowledge of Writing Javanese Script, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5417267