



Knowledge Discovery for Identifying Dominant Attributes of Student Depression Using Random Forest

Knowledge Discovery untuk Identifikasi Atribut Dominan Depresi Mahasiswa Menggunakan Random Forest

**Nur Salwa Fadia Akmar^{1*}, Gerri Asa Saputra², M. Suandi³, Muhammad Yusuf⁴,
Alsella Meiriza⁵, Ken Ditha Tania⁶, Ahmad Rifai⁷**

^{1,2,3,4,5,6,7}Program Studi Sistem Informasi, Universitas Sriwijaya, Indonesia

E-Mail: ¹09031282328066@student.unsri.ac.id, ²09031282328046@student.unsri.ac.id,
³09031282328036@student.unsri.ac.id, ⁴09031282328089@student.unsri.ac.id, ⁵allsella@unsri.ac.id,
⁶kenya.tania@gmail.com, ⁷ahmadrifai@ilkom.unsri.ac.id

Received Apr 15th 2026; Revised May 08th 2026; Accepted May 28th 2026; Available Online Jun 21th 2026

Corresponding Author: Nur Salwa Fadia Akmar

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Depression is a common mental health disorder among students, influenced by social, academic, and psychological factors. This study aims to identify dominant attributes contributing to student depression using a data mining approach within the Knowledge Discovery in Databases (KDD) framework. The novelty of this study lies in integrating Weight by Information Gain for feature selection and Random Forests for classification to quantitatively measure attribute contributions. The Student Mental Health dataset used consists of 101 instances with an imbalanced class distribution. The results show that Marital Status is the most dominant attribute, followed by Treatment, Anxiety, and Panic Attack. The model achieved an accuracy of 80.36%. This study demonstrates that the proposed approach not only classifies but also objectively identifies the dominant factors influencing student depression.

Keywords: Data Mining, Knowledge Discovery in Databases, Random Forest, Student Depression

Abstrak

Depresi merupakan gangguan kesehatan mental yang umum dialami Mahasiswa dan dipengaruhi oleh faktor sosial, akademik, serta psikologis. Penelitian ini bertujuan untuk mengidentifikasi atribut dominan depresi pada Mahasiswa menggunakan pendekatan *data mining* dalam kerangka *Knowledge Discovery in Databases* (KDD). Kebaruan penelitian ini terletak pada integrasi seleksi fitur *Weight by Information Gain* dan klasifikasi Random Forest untuk mengukur kontribusi atribut secara kuantitatif. *Dataset Student Mental Health* yang digunakan berjumlah 101 data, dengan distribusi kelas yang tidak seimbang. Hasil menunjukkan bahwa *Marital Status* merupakan atribut yang paling dominan, diikuti oleh *Treatment*, *Anxiety*, dan *Panic Attack*. Model menghasilkan akurasi sebesar 80,36%. Penelitian ini menunjukkan bahwa pendekatan yang digunakan tidak hanya mampu melakukan klasifikasi, tetapi juga mampu mengidentifikasi faktor dominan depresi pada Mahasiswa secara objektif.

Kata Kunci: Data Mining, Depresi Mahasiswa, Knowledge Discovery in Databases, Random Forest

1. PENDAHULUAN

Depresi merupakan salah satu gangguan mental yang umum dialami Mahasiswa dan memiliki dampak signifikan terhadap prestasi akademik, motivasi belajar, serta kesejahteraan individu [1]. Dibandingkan dengan gangguan kesehatan mental lainnya seperti stres dan kecemasan, depresi memiliki dampak yang lebih kompleks karena berkaitan langsung dengan penurunan fungsi kognitif, motivasi, serta risiko penurunan performa akademik yang signifikan. Selain itu, depresi sering kali tidak terdeteksi secara dini sehingga berpotensi menimbulkan konsekuensi jangka panjang terhadap keberhasilan studi Mahasiswa. Oleh karena itu, identifikasi faktor-faktor yang berkontribusi terhadap depresi Mahasiswa menjadi penting sebagai dasar deteksi awal berbasis data. Mahasiswa merupakan kelompok yang rentan mengalami depresi akibat berbagai tekanan, baik dari tuntutan akademik, lingkungan sosial, maupun faktor psikososial [2]. Untuk memahami

pola kerentanan tersebut, berbagai penelitian telah memanfaatkan pendekatan *data mining* terhadap *Dataset Student Mental Health* yang mencakup faktor akademik, demografis, dan psikologis. Meskipun jumlah atribut dalam dataset relatif terbatas, tidak semua atribut memiliki tingkat kontribusi yang sama terhadap identifikasi depresi. Beberapa atribut dapat bersifat redundan atau memiliki pengaruh yang rendah terhadap model klasifikasi. Oleh karena itu, diperlukan proses seleksi fitur untuk mengidentifikasi atribut yang paling dominan sehingga model yang dihasilkan lebih efisien, interpretatif, serta mampu memberikan informasi yang lebih bermakna. Salah satu penelitian yang dilakukan Parhusip dkk [3], menggunakan pendekatan klusterisasi K-Means untuk melakukan segmentasi kondisi kesehatan mental Mahasiswa dan menemukan bahwa status pernikahan dapat berperan sebagai faktor protektif maupun sumber tekanan. Namun, karena menggunakan pendekatan *unsupervised learning*, penelitian tersebut [3] tidak menyediakan target prediksi yang eksplisit serta tidak mampu mengukur secara kuantitatif tingkat kontribusi masing-masing variabel terhadap risiko depresi. Dengan demikian, meskipun status pernikahan diidentifikasi sebagai faktor penting, posisinya masih bersifat deskriptif dan belum tervalidasi sebagai atribut dengan kontribusi informasi tertinggi dalam konteks prediktif.

Untuk menjawab keterbatasan tersebut, penelitian ini mengadopsi pendekatan *supervised learning* melalui integrasi *Weight by Information Gain* dan Random Forest. Penggunaan operator *Weight by Information Gain* bertujuan untuk mengetahui atribut atau fitur yang paling berpengaruh dalam pembangunan model, di mana nilai yang dihasilkan berupa nilai *entropy* yang digunakan untuk mengukur reduksi entropi sehingga bobot signifikansi setiap atribut dapat ditentukan secara kuantitatif [4]. Entropi merupakan ukuran ketidakmurnian atau ketidak teraturan data, dimana nilainya berada pada range 0-1 [5]. Selanjutnya, algoritma Random Forest digunakan untuk memvalidasi signifikansi fitur tersebut dalam membangun model klasifikasi risiko depresi yang akurat. Pemilihan metode klasifikasi (*supervised learning*) ini didasari oleh keterbatasan pada penelitian terdahulu yang menggunakan pendekatan klusterisasi (*unsupervised learning*). Penelitian Parhusip dkk [3] menerapkan pendekatan klusterisasi (*unsupervised learning*) menggunakan algoritma K-Means dengan kemampuannya dalam mengelompokkan data dengan jumlah cluster yang telah ditentukan secara eksplisi [6] namun K-Means memiliki kelemahan terletak pada kendala dalam perhitungan yang relative lambat [7] sehingga kurang efektif dalam memberikan prediksi label diagnosa yang tegas pada level individu. Selain itu, metode tersebut belum menyediakan mekanisme pembobotan atribut secara eksplisit, sehingga kontribusi masing-masing variabel belum dapat diketahui secara kuantitatif. Oleh karena itu, penelitian ini menggunakan *Weight by Information Gain* untuk menyeleksi fitur yang relevan, kemudian dilanjutkan dengan klasifikasi menggunakan Random Forest. Pendekatan ini diharapkan dapat membantu dalam mengidentifikasi atribut yang lebih berpengaruh serta menghasilkan model klasifikasi yang baik.

Kontribusi ilmiah penelitian ini tidak hanya terletak pada pembangunan model klasifikasi, tetapi juga pada identifikasi atribut dominan depresi Mahasiswa secara kuantitatif. Dengan mengintegrasikan *Weight by Information Gain* dan Random Forest dalam kerangka *Knowledge Discovery*, penelitian ini memberikan pendekatan yang tidak hanya menghasilkan akurasi prediksi, tetapi juga mampu mengungkap faktor-faktor yang paling berpengaruh terhadap depresi Mahasiswa secara objektif berbasis data. Selain memberikan kontribusi pada pengembangan metode klasifikasi berbasis integrasi *Weight by Information Gain* dan Random Forest, penelitian ini juga berperan dalam kerangka *Knowledge Discovery*, yaitu proses mengekstraksi pengetahuan baru dari data [8] melalui proses pembobotan atribut dan klasifikasi, penelitian ini diharapkan dapat menghasilkan pengetahuan yang bernilai dan dapat dimanfaatkan sebagai dasar dalam memahami faktor-faktor risiko secara lebih objektif [9] dan berbasis data.

Meskipun penelitian ini telah menunjukkan bahwa integrasi metode seleksi fitur dan klasifikasi mampu memberikan hasil yang lebih interpretatif, masih terdapat kebutuhan untuk merumuskan secara lebih spesifik permasalahan penelitian yang diangkat. Oleh karena itu, rumusan masalah dalam penelitian ini difokuskan pada bagaimana mengidentifikasi atribut dominan yang berkontribusi terhadap depresi Mahasiswa serta bagaimana mengukur tingkat kontribusi masing-masing atribut secara kuantitatif dalam kerangka data mining. Selain itu, penelitian ini juga berupaya menjawab kesenjangan penelitian sebelumnya yang cenderung berfokus pada performa model tanpa memberikan interpretasi yang mendalam terhadap faktor-faktor yang mempengaruhi depresi. Dengan demikian, penelitian ini tidak hanya berorientasi pada hasil klasifikasi, tetapi juga pada pemahaman yang lebih komprehensif mengenai faktor dominan depresi Mahasiswa berbasis data.

2. METODE PENELITIAN

Penelitian ini mengimplementasikan pendekatan *supervised learning* yang merupakan metode dalam machine learning yang memanfaatkan data yang telah diberi label atau dataset yang sudah diketahui oleh pembuatnya [10]. Implementasi penelitian ini dilakukan menggunakan perangkat lunak RapidMiner atau platform perangkat lunak yang kuat untuk ilmu data dan pembelajaran mesin [11] Melalui integrasi Melalui integrasi *Weight by Information Gain* yang akan menyeleksi fitur [12] dan algoritma Random Forest yang

akan mengklasifikasikan. Tabel 1 dan 2 merupakan tabel spesifikasi perangkat lunak serta spesifikasi parameter algoritma Random Forest.

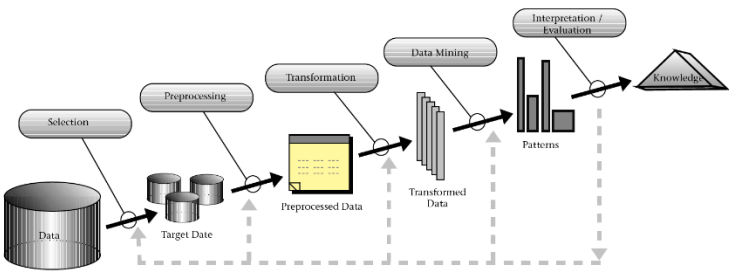
Tabel 1. Spesifikasi Perangkat Lunak dan Fungsi

Komponan	Versi/Detail	Fungsi Utama	Tahapan Penggunaan
RapidMiner Studio	2026.0.5	Tools utama seluruh proses pengolahan data	Seluruh tahapan
Weight by Information Gain	-	Menghitung bobot kepentingan setiap variable	Identifikasi atribut dengan bobot tertinggi
Random Forest	100 Trees	Memprediksi risiko depresi	Tahap pemodelan dan prediksi
Cross Validation	10-Fold	Memvalidasi performa model secara berulang	Tahap evaluasi akurasi dan presisi
Performance (Classification)	80.36%	Menampilkan metrik hasil pengujian	Tahap analisis hasil akhir model

Tabel 2. Spesifikasi Parameter Algoritma Random Forest

Parameter	Nilai	Justifikasi Teknis
<i>Number of Trees</i>	100	Titik konvergensi optimal di mana penambahan pohon tidak lagi meningkatkan akurasi secara signifikan.
<i>Criterion</i>	<i>Information Gain</i>	Menentukan pemisahan node berdasarkan pengurangan entropi paling maksimal.
<i>Maximal Depth</i>	10	Membatasi kedalaman pohon untuk mencegah <i>overfitting</i> pada dataset yang terbatas.
<i>Apply Prepruning</i>	<i>True</i>	Menghentikan pertumbuhan cabang yang tidak memiliki kontribusi statistik yang kuat.
<i>Voting Strategy</i>	<i>Confidence Vote</i>	Penentuan klasifikasi akhir berdasarkan tingkat kepercayaan dari mayoritas pohon.

Proses penelitian dilakukan dengan beberapa tahapan utama berdasarkan pendekatan *Knowledge Discovery in Databases* (KDD), yaitu pengumpulan data, pemrosesan data, transformasi data, penerapan algoritma, dan evaluasi model.[13].



Gambar 1. Alur Tahapan KDD

Penelitian ini menggunakan kerangka KDD seperti pada Gambar 1 sebagai pendekatan sistematis untuk mengekstraksi pengetahuan dari data[8] KDD terdiri dari beberapa tahapan utama, yaitu selection, preprocessing, transformation, data mining, evaluation, dan knowledge [13]. Alur proses pendekatan KDD pada penelitian ini dapat dilihat pada Gambar 2.

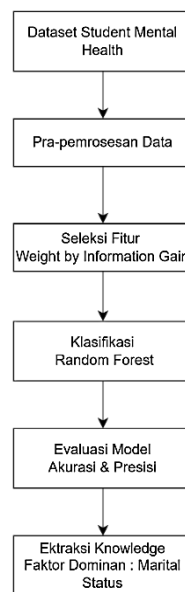
Penelitian mengenai deteksi depresi pada Mahasiswa telah banyak dilakukan seiring meningkatnya perhatian terhadap kesehatan mental di lingkungan akademik. Berbagai pendekatan telah digunakan, mulai dari metode konvensional hingga pemanfaatan teknologi *machine learning*. Penggunaan *machine learning* dinilai mampu meningkatkan akurasi deteksi serta mendukung proses identifikasi dini kondisi depresi secara lebih objektif dan efisien. Beberapa penelitian telah menerapkan berbagai algoritma *machine learning* seperti Naïve Bayes [14], K-Nearest Neighbor, Decision Tree [15], [16], Random Forest [17], dalam mengklasifikasikan kondisi depresi Mahasiswa. Hasil penelitian menunjukkan bahwa algoritma Random Forest [17] menunjukkan performa yang baik diikuti juga oleh algoritma Naïve Bayes[14] yang juga

menunjukkan performa yang cukup baik. Hal ini menunjukkan bahwa pendekatan berbasis *machine learning* memiliki potensi yang tinggi dalam meningkatkan akurasi prediksi depresi Mahasiswa.

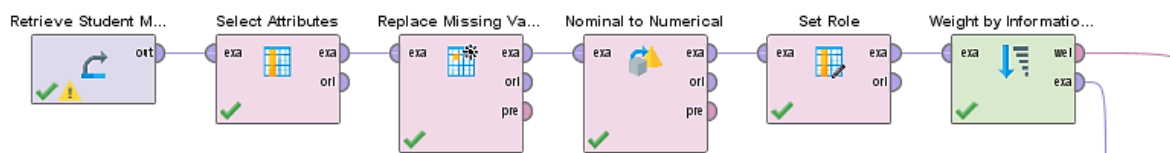
Meskipun berbagai penelitian telah berhasil meningkatkan akurasi model, sebagian besar penelitian tersebut masih berfokus pada perbandingan performa algoritma tanpa mengkaji kontribusi masing-masing atribut terhadap variabel target. Akibatnya, model yang dihasilkan cenderung bersifat “*black box*” dan kurang memberikan interpretasi mengenai faktor-faktor dominan yang memengaruhi kondisi depresi Mahasiswa. Untuk mengatasi keterbatasan tersebut, beberapa penelitian mulai mengintegrasikan teknik seleksi fitur seperti *Information Gain* untuk mengidentifikasi atribut yang paling berpengaruh dalam proses klasifikasi. Metode ini mampu mengukur tingkat kontribusi masing-masing atribut secara kuantitatif sehingga menghasilkan model yang lebih interpretatif dan efisien.

Oleh karena itu, penelitian ini mengusulkan pendekatan yang mengintegrasikan metode *Weight by Information Gain* dengan algoritma Random Forest dalam kerangka KDD. Pendekatan ini tidak hanya bertujuan untuk meningkatkan performa model klasifikasi, tetapi juga untuk mengidentifikasi atribut dominan yang berkontribusi terhadap depresi Mahasiswa secara lebih komprehensif.

Dataset yang digunakan pada penelitian ini berjumlah 101 data didapat dari platform Kaggle yang bernama “Student Mental Health” dan dipublikasikan oleh Shariful Islam. Kaggle merupakan salah satu platform populer di bidang Data Science dan Machine Learning yang menyediakan lebih dari 6.000 dataset. Dataset tersebut dapat diunduh dan umumnya tersedia dalam format CSV [18] Distribusi kelas pada variabel target menunjukkan bahwa terdapat 35 data Mahasiswa yang mengalami depresi dan 66 data tidak depresi, sehingga dataset bersifat tidak seimbang (*imbalanced dataset*). Karakteristik dataset didominasi oleh atribut kategorikal seperti *Gender*, *Course*, *Marital Status*, *Anxiety*, *Panic Attack*, dan *Treatment*, serta atribut numerik seperti *Age*. Ketidakseimbangan data ini berpotensi memengaruhi performa model, khususnya dalam mendeteksi kelas minoritas (depresi). Dengan diperolehnya data, maka penelitian dapat melanjutkan proses selanjutnya yakni *Preprocessing* yang bertujuan untuk membersihkan dan menyiapkan data sebelum dilakukann analisis [19] Pada tahapan ini digunakan 5 operator yang akan membantu proses penyiapan data sebelum dilanjutkan oleh algoritma *Random Fores*, lihat pada Gambar 3.



Gambar 2. Diagram Alur Proses Pendekatan KDD



Gambar 3. Tahap Preprocessing

Preprocessing memiliki tahapan selanjutnya yakni *Data Selection*, Dataset diambil dari *repository* RapidMiner yang sebelumnya diimpor dari sumber Kaggle, yang berisi atribut terkait kondisi mental Mahasiswa dan variabel target depresi melalui operator Retrieve. Dataset asli dari Kaggle memuat atribut *Timestamp* yang merepresentasikan waktu pengisian kuesioner, namun atribut tersebut dieliminasi melalui operator *select attributes* guna mengurangi *noise* dalam proses analisis. Setelah melalui tahap Data Selection,

maka dilakukan tahap Data Cleaning menggunakan operator Replace Missing Values untuk menangani nilai yang hilang [16] agar dataset tetap konsisten dan dapat diolah oleh algoritma klasifikasi. Tahapan selanjutnya adalah *Data Transformation*. Digunakan operator *Nominal to Numeric* untuk mentransformasikan atribut kategorikal menjadi numerik menggunakan skema Unique Integers agar dapat diproses oleh algoritma. *Operator Set Role* digunakan untuk menetapkan atribut *Depression* sebagai variabel target (label) dalam proses klasifikasi, yakni label *depression* atau depresi. Kemudian, *Weight by Information Gain* akan menghitung bobot setiap atribut sebelum dilakukan validasi pada tahap modeling menggunakan *cross-validation* yang termasuk ke dalam tahapan *data mining*. Setelah dilakukan tahap *preprocessing* dan transformasi, diperoleh dataset yang telah terstandar dan siap digunakan dalam proses seleksi fitur dan klasifikasi. Deskripsi variabel setelah *preprocessing* ditunjukkan pada Tabel 3.

Tabel 3. Deskripsi Variabel Dataset

Nama Variabel	Tipe Data Awal	Tipe Setelah Preprocessing	Peran	Deskripsi
<i>Gender</i>	Nominal	Integer (Encoding)	Prediktor	Jenis kelamin Mahasiswa
<i>Age</i>	Numerik	Numerik	Prediktor	Usia responden
<i>Course</i>	Nominal	Integer (Encoding)	Prediktor	Program studi yang diambil Mahasiswa
<i>Year of Study</i>	Ordinal	Integer (Encoding)	Prediktor	Tahun akademik Mahasiswa
<i>CGPA</i>	Ordinal	Integer (Encoding)	Prediktor	Indeks Prestasi Kumulatif Mahasiswa
<i>Marital Status</i>	Nominal	Integer (Encoding)	Prediktor	Status pernikahan Mahasiswa
<i>Depression</i>	Nominal	Integer (Label)	Target	Indikasi adanya depresi
<i>Anxiety</i>	Nominal	Integer (Encoding)	Prediktor	Indikasi adanya kecemasan
<i>Panic Attack</i>	Nominal	Integer (Encoding)	Prediktor	Riwayat serangan panik
<i>Treatment</i>	Nominal	Integer (Encoding)	Prediktor	Status pernah menjalani pengobatan

Proses terakhir adalah *data mining*. Ditahapan ini, bertujuan untuk membangun model klasifikasi menggunakan algoritma Random Forest, algoritma pembelajaran mesin dengan menghasilkan banyak pohon keputusan [21]. Proses ini dilakukan dengan metode cross-validation pada RapidMiner, di mana dataset digunakan secara bergantian sebagai data training dan testing untuk menghasilkan model klasifikasi yang optimal [22].

3. HASIL DAN PEMBAHASAN

Setelah tahap *preprocessing* dan eliminasi atribut *non-relevant*, dilakukan seleksi fitur menggunakan metode *Weight by Information Gain* untuk mengidentifikasi atribut dengan kontribusi terbesar terhadap variabel [23] depresi. Berdasarkan hasil pembobotan menggunakan *Weight by Information Gain*, diperoleh tingkat kontribusi masing-masing atribut terhadap variabel target.

attribute	weight
Choose your gender	0.015
What is your course?	0.012
Your current year of Study	0.004
What is your CGPA?	0.038
Marital status	0.286
Do you have Anxiety?	0.053
Do you have Panic attack?	0.043
Did you seek any specialist for a treatment?	0.096
Age	0.008

Gambar 8. Tabel Hasil Seleksi Fitur *Weight by Information Gain* RapidMiner

Pada Gambar 8 dan 9 hasil menunjukkan bahwa *Marital Status* memiliki bobot tertinggi sebesar 0.286. Nilai ini jauh lebih besar dibandingkan dengan atribut lainnya, yang menunjukkan bahwa status

dilanjutkan dengan mempertimbangkan faktor psikologis, serta faktor demografis yang muncul pada beberapa cabang sebagai variabel pemisah lanjutan dan juga faktor akademik yang digunakan pada level lebih dalam sebagai faktor pendukung dalam proses klasifikasi terlihat pada Tabel 5.

Tabel 5. Faktor Psikologis, Demografis, dan Akademik

Psikologis	Demografis	Akademik
<i>Anxiety</i>	Age	Course
<i>Panic Attack</i>		Year of Study
<i>Treatment</i>		CGPA

3.1. Hasil Evaluasi

Evaluasi Model dilakukan menggunakan *10-Fold Cross Validation* untuk mengukur kemampuan model Random Forest dalam mengklasifikasikan kondisi depresi Mahasiswa yang berupa *confusion matrix* pada Tabel 6 serta Hasil akhir evaluasi model.

Tabel 6. Confusion Matrix

	True Yes	True No
Pred. Yes	19	4
pred. No	16	62

TP (True Positive) : 19

TN (True Negative) : 62

FP (False Positive) : 4

FN (False Negative) : 16

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Berdasarkan *confusion matrix*, model mampu mengklasifikasikan 19 data depresi dengan benar (*True Positive*) dan 62 data tidak depresi dengan benar (*True Negative*). Namun, masih terdapat 16 data depresi yang tidak terdeteksi (*False Negative*), yang menunjukkan adanya kelemahan model dalam mengenali kelas depresi.

3.2. Hasil Perhitungan

Berdasarkan angka dari *confusion matrix* yang dihasilkan RapidMiner dan juga perhitungan rumus evaluasi, hasil akhir evaluasi model pada Tabel 7.

Tabel 7. Faktor Psikologis, Demografis, dan Akademik

	True Yes	True No	Class precision
pred. Yes	19	4	82.61%
pred. No	16	62	79.49%
Class recall	54.29%	93.94%	

Performa model Random Forest dalam mengklasifikasikan kondisi depresi Mahasiswa dapat dijelaskan secara kuantitatif. Dari 35 Mahasiswa yang benar-benar mengalami depresi, model berhasil mengidentifikasi 19 Mahasiswa, sehingga tingkat keberhasilan (*recall*) adalah 54,29%. Sementara itu, sebanyak 16 Mahasiswa depresi tidak terdeteksi, yang menunjukkan bahwa model masih melewatkan sekitar 45,71% kasus depresi yang sebenarnya. Sebaliknya, dari 66 Mahasiswa yang tidak mengalami depresi, model berhasil mengklasifikasi dengan benar 62 Mahasiswa, dengan tingkat keberhasilan tidak depresi sebesar 93,94%. Hanya 4 Mahasiswa yang salah diprediksi menderita depresi. Dari sisi ketepatan prediksi, ketika model memprediksi bahwa seorang Mahasiswa mengalami depresi, prediksi tersebut benar sebesar 82,61%, sedangkan ketika model memprediksi bahwa Mahasiswa tidak mengalami depresi, ketepatan prediksi mencapai 79,49%. Secara keseluruhan, model Random Forest menghasilkan tingkat akurasi sebesar 80,36% yang menunjukkan bahwa 80% hasil klasifikasi model sesuai dengan kondisi aktual. Meskipun model menghasilkan akurasi yang cukup tinggi yaitu 80,36%, nilai *recall* untuk kelas depresi yang hanya sebesar 54,29% menunjukkan bahwa kemampuan model dalam mendeteksi Mahasiswa yang benar-benar mengalami depresi masih belum optimal. Hal ini mengindikasikan bahwa model cenderung lebih baik dalam mengklasifikasikan data tidak depresi dibandingkan depresi. Kondisi tersebut kemungkinan dipengaruhi oleh

distribusi dataset yang tidak seimbang, sehingga model lebih dominan mengenali kelas mayoritas. Oleh karena itu, diperlukan pendekatan tambahan seperti teknik balancing data atau optimasi model untuk meningkatkan performa, khususnya pada deteksi kasus depresi.

3.3. Temuan Penelitian

Berdasarkan keseluruhan tahapan *Knowledge Discovery in Databases* (KDD) yang meliputi seleksi fitur, pemodelan menggunakan Random Forest, serta evaluasi performa model, diperoleh beberapa temuan penting terkait faktor dominan yang berkontribusi terhadap kondisi depresi Mahasiswa. Hasil seleksi fitur menggunakan *Weight by Information Gain* menunjukkan bahwa atribut *Marital Status* memiliki nilai bobot tertinggi sebesar 0.286, sehingga menjadi faktor paling informatif dalam proses klasifikasi depresi. Temuan ini diperkuat oleh struktur model Random Forest, di mana atribut tersebut muncul sebagai node akar pada salah satu pohon keputusan, yang menunjukkan perannya sebagai faktor awal dalam proses pemisahan data. Selain itu, atribut yang berkaitan dengan kondisi kesehatan mental, seperti *Treatment*, *Anxiety*, dan *Panic Attack*, juga menunjukkan kontribusi yang signifikan dalam mendukung proses klasifikasi. Sebaliknya, atribut akademik seperti *CGPA*, *Course*, *Year of Study* memiliki bobot yang relatif lebih rendah, sehingga berperan sebagai faktor pendukung dibandingkan dengan faktor utama. Temuan ini mengindikasikan bahwa kondisi depresi Mahasiswa lebih erat kaitannya dengan faktor sosial dan psikologis dibandingkan dengan faktor akademik hal ini didukung salah satu penelitian yang dilakukan, penelitian terdahulu [25] yang mengatakan faktor akademik juga masih menyesuaikan tingkatannya terhadap depresi Mahasiswa. [25].

Hasil evaluasi model juga menunjukkan bahwa meskipun model memiliki kemampuan yang sangat baik dalam mengenali Mahasiswa yang tidak mengalami depresi, deteksi terhadap Mahasiswa yang mengalami depresi masih lebih terbatas. Hal ini memperkuat pentingnya faktor-faktor dominan yang telah diidentifikasi sebagai indikator utama dalam klasifikasi. Implikasi dari penelitian ini menunjukkan bahwa pendekatan berbasis *data mining* dapat digunakan sebagai alat bantu dalam mendukung deteksi dini kondisi depresi Mahasiswa. Informasi mengenai atribut dominan yang dihasilkan, seperti status pernikahan dan faktor psikologis, dapat dimanfaatkan oleh institusi pendidikan dalam merancang kebijakan atau program intervensi yang lebih tepat sasaran. Selain itu, hasil penelitian ini juga dapat menjadi dasar dalam pengembangan sistem pendukung keputusan berbasis data untuk membantu pihak kampus dalam mengidentifikasi Mahasiswa yang berisiko mengalami depresi secara lebih objektif.

4. KESIMPULAN

Berdasarkan hasil penelitian melalui tahapan *Knowledge Discovery in Database* (KDD), metode *Weight by Information Gain* dan Random Forest terbukti mampu mengidentifikasi atribut dominan dalam klasifikasi depresi Mahasiswa dengan tingkat akurasi sebesar 80,36%. Hasil *data mining* menunjukkan bahwa *Marital Status* menjadi faktor paling berpengaruh, diikuti oleh faktor psikologis seperti *Treatment*, *Anxiety*, dan *Panic Attack*, sementara faktor akademik memiliki kontribusi yang lebih rendah sehingga diperoleh pengetahuan bahwa kondisi depresi Mahasiswa lebih dipengaruhi oleh aspek sosial dan kesehatan mental dibandingkan dengan faktor akademik. Oleh karena itu, penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar dan seimbang serta menerapkan teknik penanganan data tidak seimbang seperti SMOTE atau undersampling guna meningkatkan kemampuan model dalam mendeteksi kelas minoritas. Selain itu, penggunaan metode klasifikasi lain atau pendekatan ensemble yang lebih kompleks dapat dibandingkan untuk memperoleh performa yang lebih optimal, khususnya dalam meningkatkan nilai *recall* pada kasus depresi Mahasiswa.

REFERENSI

- [1] N. N. M. I. AA Zubairi, "The Impact of Between Depression, Anxiety, And Stress on Academic Performance among University Students," *Jurnal Pendidikan Bitara UPSI*, no. 18, pp. 22–29, May 2025, doi: 10.37134/bitara.vol17.1.3.2025.
- [2] I. Lazulfa, Z. Atsna, and N. Angelia, "Analisis Hubungan Stress Akademik Dengan Gejala Depresi Dan Ide Bunuh Diri Pada Mahasiswa," *Journal of Telenursing (JOTING)*, vol. 7, no. 3, p. 2025, doi: 10.31539/joting.v7i3.15445.
- [3] J. Parhusip, J. Feranita, M. Pradana Mentari, T. Apriliani, and A. Marcello Jessend, "Implementasi Algoritma K-Means Clustering Untuk Segmentasi Tingkat Kesehatan Mental Mahasiswa Berdasarkan Faktor Stress dan Akademik," 2026.
- [4] E. O. Ali, M. Agustin, and R. Sari, "Implementasi Algoritma Decision Tree Dengan Fitur Seleksi Weight By Information Gain," *Multinetics*, vol. 9, no. 2, pp. 118–126, 2024, doi: 10.32722/multinetics.v9i2.5715.
- [5] A. D. W. M. Sidik, I. Himawan Kusumah, A. Suryana, Edwinanto, M. Artiyasa, and A. Pradiftha Junfithrana, "Gambaran Umum Metode Klasifikasi Data Mining," *FIDELITY: Jurnal Teknik Elektro*, vol. 2, no. 2, pp. 34–38, 2020, doi: 10.52005/fidelity.v2i2.111.

- [6] A. F. N. Masruriyah, M. Mardiah, M. D. Ananda, and K. N. Malik, "Pendekatan *Unsupervised learning* dalam Segmentasi Kesehatan: Perbandingan K-Means dan DBSCAN," *MIND Journal*, vol. 10, no. 1, pp. 99–113, Jun. 2025, doi: 10.26760/mindjournal.v10i1.99-113.
- [7] B. Nurhakim, I. Septiani, K. Anam, and D. Pratama, "Penerapan Algoritma K-Means Clustering Dalam Menganalisis Resiko Penyakit Stroke," 2024.
- [8] P. Ayuningtiyas, K. Ditha Tania, and W. Kurnia Sari, "Sentiment-Based Knowledge Discovery pada Aplikasi iPusnas Menggunakan Metode Machine Learning dan Deep Learning," 2025. [Online]. Available: <http://jurnal.polibatam.ac.id/index.php/JAIC>
- [9] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," 1996. [Online]. Available: www.aaai.org
- [10] Z. N. Syarif, "Teknois : Jurnal Ilmiah Teknologi Informasi dan Sains [1] Penerapan Information Gain dan Algoritma K-Means Untuk Klasterisasi Kedisiplinan Pegawai Menggunakan Rapidminer," vol. 13, no. 1, pp. 1–12, 2023, doi: 10.36350/jbs.v13i1.
- [11] M. R. Nahjan, N. Heryana, and A. Voutama, "Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko Oj Cell," 2023.
- [12] A. Budianita, "Information Gain Berbasis Algoritma Naive Bayes Classifier Pada Pemodelan Prediksi Kelulusan," *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 5, no. 1, pp. 1–10, May 2023, doi: 10.46772/intech.v5i1.1116.
- [13] D. Athallah Ubaid and M. Ichsan Farel Rachmad, "Klasifikasi Risiko Strok Menggunakan Algoritma Random Forest dengan Teknik Knowledge Discovery in Database," *Journal Computer Science*, vol. 4, no. 1, 2025, [Online]. Available: <https://www.kaggle.com/datasets/mahatiratushe>
- [14] N. Supriani, S. W. Ningrum, and A. G. Nugraheni, "Technology and Informatics Insight Journal Penerapan Algoritma Naive Bayes dalam Machine Learning Untuk Klasifikasi Faktor-Faktor yang Mempengaruhi Kesehatan Mental Siswa", [Online]. Available: <https://jurnal.universitaspurabangsa.ac.id/index.php/tiij>
- [15] S. Zacksavira, C. Anggraeni, B. Sebayang, J. Imanuel, and G. B. Sibarani, "Copyright @ Implementasi Algoritma Decision Tree untuk Deteksi Depresi Pada Pelajar".
- [16] J. Sarita, I. Lu'luatuz Zainiyah, I. Maylani, and P. Setyawati, "Pengaruh Kesehatan Mental Terhadap Stress Akademik Pada Mahasiswa Di Universitas X Dengan Menggunakan Metode Algoritma Decision Tree J48."
- [17] A. Y. Pratama, I. S. Maulana, F. K. Sari, S. D. Tiara, and I. Darmawan, "Prediksi Risiko Depresi pada Mahasiswa Menggunakan Algoritma Random Forest Berdasarkan Data Akademik dan Gaya Hidup," *JSITIK: Jurnal Sistem Informasi dan Teknologi Informasi Komputer*, vol. 4, no. 1, pp. 1–10, Dec. 2025, doi: 10.53624/jsitik.v4i1.696.
- [18] A. Rahmat, M. Syafiih, and M. Faid, "Implementasi Klasifikasi Potensi Penyakit Jantung Dengan Menggunakan Metode C4.5 Berbasis Website (Studi Kasus Kaggle.Com)," *Infotech*, vol. 1861, no. 9, pp. 393–400, 2023.
- [19] F. N. Hidayat and S. Sugiyono, "Analisis Sentimen Masyarakat Terhadap Perekrutan Pppk Pada Twitter Dengan Metode Naive Bayes Dan Support Vector Machine," *Jurnal Sains dan Teknologi*, vol. 5, no. 2, pp. 665–672, Dec. 2023, doi: 10.55338/saintek.v5i2.1359.
- [20] L. N. Darussalam, R. Kurniawan, Y. A. Wijaya, and T. Suprpti, "Algoritma Naive Bayes Untuk Meningkatkan Model Klasifikasi Penerima Program Indonesia Pintar Di Sdn 2 Purwawinangun," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 1, pp. 2830–7062, doi: 10.23960/jitet.v13i1.5882.
- [21] M. M. Mutoffar *et al.*, "Klasifikasi Kualitas Air Sumur Menggunakan Algoritma Random Forest," vol. 04, 2022.
- [22] A. N. Yuliarina and H. Hendry, "Comparison Of Prediction Analysis Of Gofood Service Users Using The Knn & Naive Bayes Algorithm With Rapidminer Software," *Jurnal Teknik Informatika (Jutif)*, vol. 3, no. 4, pp. 847–856, Aug. 2022, doi: 10.20884/1.jutif.2022.3.4.294.
- [23] E. Sufarnap and S. Sudarto, "Seleksi Fitur Information Gain Dalam Meningkatkan Kinerja Naive Bayes Dalam Prediksi Performa Mahasiswa," *Digital Transformation Technology*, vol. 4, no. 2, pp. 1285–1295, Feb. 2025, doi: 10.47709/digitech.v4i2.5525.
- [24] M. Nurul Hamidah, R. Febrifanyang Tias, and R. Fahrial Zainal, "Analisis Komparatif Metode K-Nearest Neighbors Dan Random Forest Untuk Rekomendasi Pemilihan Pesantren Berdasarkan Profil Minat Siswa (siswa xxx tingkat SD dan SMP) A Comparative Analysis of K-Nearest Neighbors and Random Forest Methods for Recommendations on Selecting Islamic Boarding Schools Based on Student Interest Profiles (primary and middle school students at xxx)."
- [25] A. Muttaqin, U. K. Abdurrahman Wahid Pekalongan, and M. Hanif, "Pengaruh Stres Akademik Terhadap Tingkat Kecemasan dan Depresi Mahasiswa Perguruan Tinggi."