JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8871
Hal 636-647
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Klasifikasi Kesehatan Mental Mahasiswa Menggunakan Light Gradient Boosting Machine Dan Analisa Fitur Menggunakan SHAP

Ditto Ridhwan Wibowo*, Fajri Rakhmat Umbara, Ridwan Ilyas

Sains dan Informatika, Informatika, Universitas Jenderal Achmad Yani, Cimahi, Indonesia
Email: [1*] dittoridhwan21@if.unjani.ac.id, [2] fajri.rakhmat@lecture.unjani.ac.id, [3] ilyas@lecture.ac.id
Email Penulis Korespondensi: dittoridhwan21@if.unjani.ac.id
Submitted **18-06-2025;** Accepted 0**1-08-2025;** Published **30-08-2025**

Abstrak

Kesehatan mental mahasiswa menjadi isu penting karena banyak yang tidak menerima perawatan meskipun membutuhkannya. Berdasarkan Asosiasi Direktur Pusat Konseling Universitas dan Perguruan Tinggi 95% mahasiswa mengalami peningkatan pada psikopatologi. Penelitian ini menggunakan algoritma Light Gradient Boosting Machine untuk mengklasifikasikan kesehatan mental mahasiswa berdasarkan dataset yang memiliki jumlah 61.794 baris dan 16 kolom. Light Gradient Boosting Machine merupakan implementasi dari Gradient Boosting Decision Tree yang memiliki dua strategi yaitu gradient-based one-side sampling (GOSS) dan leaf-wise growth. Hasil akurasi yang diperoleh dengan menggunakan LightGBM mencapai 67% dimana data yang digunakan sudah di balancing menggunakan parameter class_weight dan teknik ADASYN. Selain itu, pada penelitian dilakukan analisa untuk mencari fitur yang paling berkontribusi dengan menggunakan metode SHAP (SHapley Additive exPlanations) dengan hasil yang diperoleh terdapat 6 fitur yang memiliki nilai kontribusi tertinggi antara lain Country, treatment, mental_health_interview, family_history, Gender, dan self_employed.

Kata Kunci: Kesehatan Mental; LightGBM; SHAP; Hyperparameter Tuning; ADASYN

Abstract

The mental health of college students is an important issue as many do not receive treatment despite needing it. According to the Association of University and College Counseling Center Directors 95% of college students experience an increase in psychopathology. This study uses the Light Gradient Boosting Machine algorithm to classify the mental health of college students based on a dataset that has a total of 61.794 rows and 16 columns. Light Gradient Boosting Machine is an implementation of Gradient Boosting Decision Tree which has two strategies namely gradient-base one-side sampling (GOSS) and leaf-wise growth. The accuracy results obtained using LightGBM reached 67% where the data used had been balanced using the class_weight parameter and the ADASYN technique. In addition, the research was analyzed to find the most contributing features using the SHAP (SHapley Additive exPlanations) method with the results obtained there are 6 features that have the highest contribution value including Country, treatment, mental_health_interview, family_history, Gender, dan self_employed.

Keywords: Mental Health; LightGBM; SHAP; Hyperparameter Tuning; ADASYN

1. PENDAHULUAN

Kesehatan mental merupakan suatu kondisi kompleks yang melibatkan berbagai aspek kehidupan, mulai dari biologis hingga sosial. Penyakit mental ini umum terjadi dan banyak orang dewasa tidak menerima perawatan kesehatan mental, meskipun perawatan tersedia. Begitupun mahasiswa, banyak mahasiswa yang melaporkan memiliki masalah kesehatan mental dan hanya sekitar sepertiga dari mahasiswa yang menerima perawatan kesehatan mental [1]. Menurut Asosisasi Direktur Pusat Konseling Universitas dan Perguruan Tinggi, 95% direktur pusat konseling melaporkan adanya peningkatan tingkat keparahan pada psikopatologi mahasiswa. Meningkatnya tingkat tekanan psikologis di kalangan mahasiswa tidak hanya terjadi di Amerika Serikat, hal serupa terjadi juga di Inggris, Australia, Selandia Baru, dan Kanada [2]. Pada penelitian yang relevan juga menemukan bahwa titik global depresi adalah 12.9% dimana Amerika Selatan mencapai 20.6%, Asia 16.7%, Amerika Utara 13.4%, Eropa 11.9%, Afrika 11.5% dan Australia 7.3% [3]. Oleh karena itu, penting untuk mengidentifikasi mahasiswa yang membutuhkan perawatan terhadap gangguan kesehatan mental. Dalam perkembangan teknologi saat ini memberikan peluang untuk mengolah sebuah data yang kompleks dan menghasilkan klasifikasi atau prediksi terkait kondisi kesehatan mental menggunakan algoritma machine learning. Salah satu algoritma yang digunakan adalah Light Gradient Boosting Machine atau yang dikenal dengan LightGBM. LightGBM adalah algoritma berbasis decision tree yang membagi parameter di lapisan input menjadi beberapa bagian [4].

Pada penelitian [5] melakukan klasifikasi kesehatan mental menggunakan metode J48 (C4.5), *Random Forest*, dan *Random Tree* dengan melakukan metrik evaluasi seperti *ROC Curve, precision, recall* dan *F-measure*. Hasil analisis menunjukkan bahwa pendekatan *Random Tree* memberikan hasil yang lebih baik dalam mengidentifikasi gejala dan kondisi kesehatan mental. Penelitian lebih lanjut disarankan menggunakan dataset dari sumber lain dengan memanfaatkan teknologi machine learning yang lebih canggih untuk meningkatkan akurasi dan relevansi model klasifikasi dalam konteks kesehatan mental.

Pada penelitian klasifikasi diabetes melitus gestasional dilakukan dengan menggunakan Random Forest dan LightGBM, dari hasil penelitian yang dilakukan akurasi yang diperoleh cukup baik namun precision yang dihasilkan tidak mencapai 50%. Celah yang dapat diambil adalah penanganan pada dataset yang tidak seimbang pada label multikelas [6]. Pada penelitian terdahulu melakukan pengembangan dan evaluasi klasifikasi menggunakan pendekatan ensemble resampling yang menggabungkan SVM dan Multinomial Regression dengan melakukan eksplorasi teknik resampling dengan menggunakan ADASYN. Hasil pada penelitian tersebut menunjukkan resampling menggunakan ADASYN

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8871
Hal 636-647
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

berpengaruh terhadap nilai sensitivitas (recall), namun pengaruh pada akurasi tidak selalu konsisten dan terkadang mengalami penurunan. Celah yang dapat diambil adalah penggunaan ADASYN dengan menggunakan model machine learning lain [7].

Penelitian [8] melakukan perbandingan metode antara *Gradient Boosting* dengan *Light Gradient Boosting* dalam mengklasifikasikan rumah sewa. Dataset yang digunakan pada penelitian tersebut terdapat 4.745 data, dimana dengan jumlah data tersebut biasanya sering terjadi *overfitting* terutama pada model yang sangat kompleks seperti *Gradient Boosting*. Celah yang dapat diambil yaitu melakukan optimasi *hyperparameter* untuk menemukan kombinasi yang tepat sehingga mencegah terjadinya *overfitting*, kemudian memperluas analisis fitur untuk memahami fitur mana yang paling berkontribusi seperti *Feature Importance* atau SHAP.

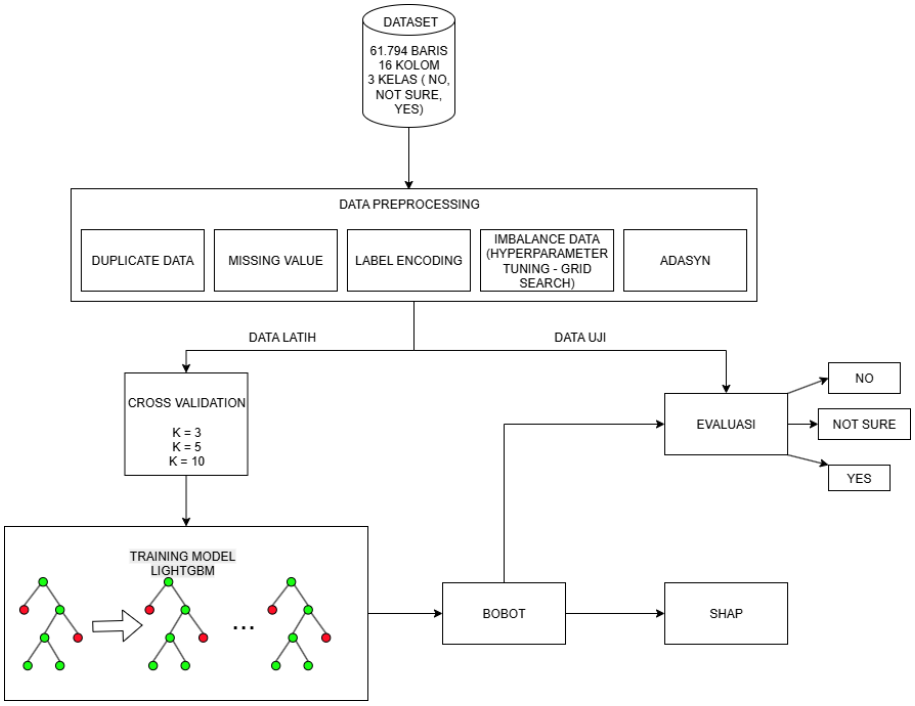
Terdapat juga pada penelitian [9] yang berfokus pada pengembangan model prediksi kesehatan mental menggunakan algoritma klasifikasi seperti Random Forest, Naïve Bayes dan Decision Tree. Namun, pendekatan pada penelitian ini memiliki keterbatasan dalam hal akurasi dan kurangnya eksplorasi data multimodal sehingga penting untuk mengeksplorasi algoritma yang lebih maju. Dataset yang digunakan diambil melalui kuisioner dimana rentan terhadap bias subyektif responden, sehingga perlu diperhatikan kembali dataset yang digunakan.

Light Gradient Boosting Machine (LightGBM) adalah metode *gradient boosting* berbasis *decision tree* yang dikenal karena kecepatan, efisiensi dan kinerjanya yang tinggi. Teknologi ini memiliki beberapa keunggulan, seperti proses pelatihan yang cepat, penggunaan memori yang rendah, kemampuan menangani dataset berukuran besar, serta menghasilkan akurasi prediksi yang lebih baik. LightGBM menggunakan pendekatan *Gradient Boosting Machine Decision Tree* yang ditingkatkan contohnya seperti *leaf-wise growth* untuk meningkatkan akurasi model secara optimal [10].

Penelitian ini diupayakan dapat mengimplementasikan algoritma LightGBM dalam klasifikasi kesehatan mental mahasiswa untuk menentukan perawatan yang diperlukan mahasiswa dalam masalah kesehatan mental. Penulis mengambil dataset dari penelitian terdahulu yang memiliki total 61.794 data dan 16 fitur [11]. Berdasarkan uraian diatas, maka peneliti akan melakukan beberapa tahap *preprocessing* seperti penghapusan data duplikasi, penanganan pada *missing value*, konversi data menjadi numerik menggunakan *Label encoding*, menangani data yang tidak seimbang menggunakan parameter *class weight* dan optimasi *hyperparameter tuning (Grid Search)*, melakukan oversampling dengan ADASYN dan terakhir metode SHAP (*SHapley Additive exPlanations*) sebagai teknik untuk menganalisis fitur mana yang paling berkontribusi dalam hasil klasifikasi.

2. METODOLOGI PENELITIAN

Berdasarkan pada Gambar 1 akan dilakukan beberapa tahap penelitian dalam klasifikasi kesehatan mental mahasiswa ini. Dataset yang diperoleh pada penelitian ini diambil dari penelitian [11] bersumber dari website Kaggle. Dataset ini berjumlah 61.794 baris yaitu, 24.993 data No, 20.304 data Yes, dan 16.497 data Not Sure. Pada data ini terdapat 16 kolom yaitu 15 kolom fitur dan 1 kolom target (care_options). Berikut merupakan alur peneletian yang dilakukan berdasarkan pada gambar 1:



Gambar 1. Alur Penelitian

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8871
Hal 636-647
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

3.1 2.2 Preprocessing

Preprocessing adalah tahapan yang bertujuan untuk membersihkan dataset yang akan digunakan dalam proses klasifikasi antara lain:

2.2.1 Handling Duplicate Data

Handling Duplicate Data adalah tahapan dalam proses pembersihan data dengan menghapus data yang duplikasi, karena jika dalam dataset terdapat data yang duplikat maka dapat menyebabkan bias dalam analisis dan dapat mempengaruhi hasil akurasi model *machine learning*.

2.2.2 Handling Missing Value

Handling missing value merupakan tahapan untuk menangani data yang hilang dari dataset. Pada penelitian ini, untuk menangani *missing value* tersebut yaitu dengan melakukan imputasi data menggunakan jumlah frekuensi terbanyak atau modus, hal ini akan digunakan pada atribut *self_employed* karena data tersebut berbentuk kategorikal dan dapat mengambil nilai-nilai tertentu dan terbatas (variabel diskrit) [12].

2.2.3 Label Encoding

Label encoding merupakan teknik memberikan nilai numerik ke setiap kategori. Metode ini sangat berguna ketika variabel kategorikal memiliki urutan tertentu tetapi juga dapat digunakan untuk variabel nominal [11]. Singkatnya, Teknik ini akan mengonversi data teks secara langsung menjadi nilai integer yang memiliki makna nominal tanpa memperhatikan urutan atau tingkatan [13]. Pada penelitian ini teknik yang digunakan adalah *label encoding* dimana nanti setiap fitur yang berisi data kategorikal akan diubah ke dalam bentuk numerik, misal pada kolom *Mood_Swings* terdapat 3 kategori yaitu, *Low, Medium* dan *High*. Dengan menggunakan teknik *label encoding*, ketiga kategori tersebut akan ditransform ke bentuk numerik menjadi 0, 1, dan 2.

2.2.4 Imbalance Data (class weight dan Hyperparameter Tuning)

Dalam menangani data yang tidak seimbang akan menggunakan *hyperparameter tuning* yang dimiliki oleh metode LightGBM dengan memanfaatkan parameter yang baik dalam menangani data tidak seimbang contohnya seperti parameter *class_weight* dan menggunakan *grid search* untuk melihat semua kombinasi parameter dalam suatu pencarian. Dengan melakukan penanganan *imbalance* data menggunakan *hyperparameter tuning* dapat membantu metode LightGBM dalam menangani bias data dan mengoptimalkan performa LightGBM dengan menggunakan persamaan 1.

$$class\ weight_i = \frac{N}{K\ x\ n_i} \qquad (1)$$

Keterangan :
N : Total jumlah sampel (semua kelas)
K : Jumlah kelas unik
n_i : Jumlah sampel pada kelas ke-i
$class\ weight_i$: Bobot untuk kelas ke-i

2.2.5 ADASYN

Kemudian setelah melakukan *balancing* dengan menggunakan *class_weight*, akan dilakukan *oversampling* dengan menggunakan metode ADASYN pada label yang termasuk dalam kategori minoritas dengan menambahkan data sintetis. ADASYN adalah versi perbaikan *Sintetis Minoritas Over-Sampling Technique* (SMOTE) yang digunakan untuk menghindari *overfitting* yang terjadi ketika replika yang tepat dari instance minoritas ditambahkan ke dataset utama [14]. ADASYN merupakan teknik untuk mengatasi masalah data *imbalance* dengan melakukan *oversampling* pada kelas minoritas dengan menggunakan bobot distribusi untuk pada kelas minoritas berdasarkan pada tingkat kesulitan pembelajaran model. Data sintetis dihasilkan dari data minoritas yang sulit untuk dipahami dibandingkan dengan data minoritas yang lebih mudah untuk dipahami [15]. Lalu dataset akan dilakukan pembagian antara data latih dan data uji dengan perbandingan 80:20.

3.2 2.3 Cross Validation

Cross Validation merupakan teknik validasi yang digunakan untuk mengevaluasi hasil analisis. Metode *K-Fold Cross Validation* digunakan untuk memperkirakan kesalahan prediksi dan mengevaluasi model dengan melakukan beberapa pengujian pada data yang diminta [16]. Pada tahap *cross validation* dilakukan tiga percobaan evaluasi dengan menggunakan data latih. Tiga percobaan ini dibagi menjadi 3 fold, 5 fold dan 10 fold dengan hasil akhirnya menggunakan F1-score, dapat dilihat pada tabel 1, tabel 2 dan tabel 3.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8871
Hal 636-647
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tabel 1. Cross Validation 3 fold

6.375	6.375	6.375	Fold 1
6.375	6.375	6.375	Fold 2
6.375	6.375	6.375	Fold 3

Tabel 2. Cross Validation 5 fold

3.825	3.825	3.825	3.825	3.825	Fold 1
3.825	3.825	3.825	3.825	3.825	Fold 2
3.825	3.825	3.825	3.825	3.825	Fold 3
3.825	3.825	3.825	3.825	3.825	Fold 4
3.825	3.825	3.825	3.825	3.825	Fold 5

Tabel 3. Cross Validation 10 Fold

1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	Fold 1
1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	Fold 2
1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	Fold 3
1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	Fold 4
1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	Fold 5
1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	Fold 6
1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	Fold 7
1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	Fold 8
1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	Fold 9
1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	1.912	Fold 10

Pada tabel 1, tabel 2 dan tabel 3 merupakan contoh ilustrasi bagaimana proses evaluasi dengan *cross validation*. Untuk kolom berwarna merah sebagai validasinya sedangkan kolom hijau sebagai trainingnya. Pada cross validation 3 fold data latih yang berjumlah 19.125 akan dibagi ke dalam 3 bagian, sehingga masing-masing fold memiliki 6.375 tiap barisnya. Contohnya pada fold 1 kolom 6.375 pertama akan digunakan sebagai data uji dan untuk kolom kedua dan ketiga digunakan sebagai data latih. Kemudian untuk cross validation 5 fold data akan dibagi menjadi 5 bagian dengan masing-masing kolom berjumlah 3.825 baris, dan untuk cross validation 10 fold data akan dibagi menjadi 10 bagian dengan masing-masing kolom berjumlah 1.912 baris. Pada tiap fold akan dihitung nilai f1-scorenya kemudian dijumlahkan untuk keseluruhan nilai foldnya dan terakhir akan dihitung nilai rata-ratanya.

3.3 2.4 LightGBM

Setelah melakukan balancing pada dataset akan dilakukan pembagian antara data latih dan data uji dengan perbandingan 80:20, kemudian untuk data latih akan dilakukan pelatihan menggunakan LightGBM. Light Gradient Boosting Machine merupakan implementasi dari Gradient Boosting Decision Tree yang memiliki dua strategi yaitu gradient-based one-side sampling (GOSS) dan leaf-wise growth [10]. Leaf-wise growth adalah teknik yang berfungsi dalam membatasi kedalaman model. Proses ini bekerja dengan mencari node yang memiliki keuntungan pemisahan (splitting gain) terbesar, kemudian memecah node tersebut dan melanjutkan ke node baru. Leaf-wise growth merupakan strategi efisien untuk menumbuhkan trees(pohon) yang dapat mengurangi lebih banyak kesalahan dan mendapatkan akurasi yang lebih baik dalam waktu pemisahan yang sama [17]. LightGBM dapat dengan cepat memproses data yang sangat besar dengan mendorong leaf-wise growth dengan batasan kedalaman dan pengambilan sampel satu sisi berbasis gradien serta pendekatan bundling fitur eksklusif [18]. Metode LightGBM melatih model dengan T trees dengan menerapkan proses pelatihan aditif dimana setiap model baru belajar untuk memprediksi model sebelumnya. Untuk membangun model LightGBM dengan T trees dapat dilakukan dengan menerapkan persamaan 2 sebagai berikut:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(X_i) \qquad (2)$$

Setiap iterasi, model y^_i tetap digunakan dan fungsi baru f atau residual yang telah dilatih ditambahkan ke dalam model [19].

3.4 2.5 Evaluasi

Tahapan evaluasi dilakukan untuk mengukur kemampuan model baik pada training set maupun test set dengan menggunakan confusion matrix yang terdiri dari beberapa komponen yang akan dievaluasi, yaitu TP (True Positive) data positif yang diprediksi dengan benar, TN (True Negative) data negatif yang diprediksi dengan benar, FP (False Positive) data negative yang terdeteksi data positif dan FN (False Negative) data positif namun terdeteksi sebagai data negatif [19]. Hasil akhir evaluasi yang akan dilakukan adalah menghitung hasil akurasi, recall, precision dan F1 – Score berdasarkan persamaan berikut:

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8871
Hal 636-647
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{F1- Score} = 2 \times \frac{precision \times recall}{precision+recall} \quad (6)$$

3.5 2.6 SHAP

SHAP atau dikenal dengan (SHapley Additive exPlanations) merupakan metode yang menggunakan nilai Shapley dari Lloyd Stowell Shapley dimana metode ini mengadopsi nilai Shapley dan independensi antar fitur dan merupakan metode numerik yang dapat menghitung kontribusi dari setiap fitur untuk menghasilkan hasil keseluruhan [20]. SHAP menyediakan pendekatan kuat untuk mengungkap kontribusi setiap fitur terhadap hasil prediksi [21]. Teknik SHAP ini menghitung kontribusi setiap fitur secara adil dengan mempertimbangkan semua kombinasi fitur lainnya. Hasil dari persamaan (7) ini akan menghasilkan nilai kontribusi dari fitur yang digunakan.

$$\phi_i(v) = \sum_{S \in N\{i\}} \frac{|S|!\,(n-|S|-1)!}{n!}\,(v(S \cup \{i\}) - v(S)) \quad (7)$$

Keterangan:
ϕ_i (v) = nilai kontribusi untuk fitur ke-i
n = jumlah semua fitur
S = subset dari semua fitur kecuali fitur ke-i dalam seluruh kelompok
|S| = jumlah elemen dalam S
v(S) = kontribusi dari subset yang tersisa tidak masuk fitur ke -i
(v(S∪{i}) = total kontribusi termasuk fitur ke-i

3. HASIL DAN PEMBAHASAN

3.6 Dataset

Berdasarkan uraian pada metode penelitian sebelumnya, penulis memperoleh dataset dari penelitian terdahulu yang diambil dari website Kaggle. Berikut merupakan fitur dari dataset yang akan digunakan berdasarkan pada tabel 4:

Tabel 4 Fitur-fitur

No	Atribut	Keterangan
1	Gender	Jenis kelamin responden.
2	Country	Negara tempat tinggal responden.
3	Occupation	Jenis pekerjaan responden.
4	self_employed	Apakah responden bekerja mandiri atau tidak.
5	family_history	Riwayat kesehatan mental dalam keluarga responden
6	treatment	Apakah responden pernah atau sedang menjalani pengobatan kesehatan mental.
7	Days_Indoors	Perkiraan jumlah hari yang dihabiskan di dalam ruangan.
8	Growing_Stress	Apakah stres meningkat selama waktu tertentu.
9	Changes_Habits	Adanya perubahan kebiasaan responden.
10	Mental_Health_History	Riwayat kesehatan mental pribadi.
11	Mood_Swings	Apakah responden mengalami perubahan suasana hati.
12	Coping_Struggles	Kesulitan dalam menangani masalah atau beban.
13	Work_Interest	Minat terhadap pekerjaan
14	Social_Weakness	Kelemahan dalam interaksi sosial.
15	mental_health_interview	Kesediaan untuk membahas kesehatan mental dalam wawancara.
16	care_options	Opsi perawatan kesehatan mental untuk responden.

Pada penjelasan sebelumnya pada dataset ini terdapat 61.794 baris dengan 16 kolom antara lain, Gender, Country, Occupation, self_employed, family_history, treatment, Days_Indoors, Growing_Stress, Changes_Habits, Mental_Health_History, Mood_Swings, Coping_Struggles, Work_Interest, Social_Weakness, mental_health_interview, dan care_options. Dengan kolom care_options sebagai kolom target yang memiliki 3 kelas yaitu, No, Yes dan Not Sure.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8871
Hal 636-647
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

3.7 Preprocessing

3.2.1 Handling Duplicate Data

Dalam penanganan data duplikasi, penulis melakukan pemeriksaan untuk melihat jumlah data duplikasi pada dataset tersebut, hasilnya dapat dilihat pada Gambar 2 dimana diperoleh duplikasi pada dataset dengan jumlah 41.048 baris untuk itu penanganan yang dilakukan adalah menghapus data duplikasi tersebut agar saat model melatih data tersebut tidak akan menyebabkan bias data atau adanya overfitting pada model.

61787	Male	Australia	Student	No	Yes	Yes	More than 2 months
61788	Male	United States	Student	No	No	No	More than 2 months
61791	Male	United States	Student	No	Yes	No	More than 2 months
61792	Male	United States	Student	No	Yes	Yes	More than 2 months
61793	Male	United States	Student	No	Yes	Yes	More than 2 months

41048 rows × 16 columns

Gambar 2. Cek Duplikasi Data

3.2.2 Handling Missing Value

Dalam penanganan data yang hilang juga penulis melakukan pemeriksaan untuk melihat jumlah data yang hilang berdasarkan dataset tersebut, hasilnya dapat dilihat pada Gambar 3

```
Gender                     0
Country                    0
Occupation                 0
self_employed           1107
family_history             0
treatment                  0
Days_Indoors               0
Growing_Stress             0
Changes_Habits             0
Mental_Health_History      0
Mood_Swings                0
Coping_Struggles           0
Work_Interest              0
Social_Weakness            0
mental_health_interview    0
care_options               0
dtype: int64
```

Gambar 3. Cek Nilai Hilang

Berdasarkan pada Gambar 3 dari 16 fitur terdapat 1 fitur yang memiliki data yang hilang yaitu pada kolom self_employed sebanyak 1.107 data. Untuk menangani hal tersebut penulis melakukan imputasi data dengan menggunakan nilai modus atau berdasarkan kategori yang paling banyak pada kolom self_employed. Hal tersebut dilakukan karena jika data yang hilang tersebut dihapus dapat menyebabkan kurangnya informasi pada saat model LightGBM melakukan training data. Sehingga jumlah data setelah melakukan proses pembersihan data adalah 20.746, data tersebut yang akan digunakan pada penelitian klasifikasi kesehatan mental mahasiswa berdasarkan pada Gambar 4.

Gender	Country	Occupation	self_employed	family_history	treatment	Days_Indoors	Growing_Stress	Changes_Habits	Mental_Health_History
20746	20746	20746	20746	20746	20746	20746	20746	20746	20746
2	35	1	2	2	2	5	3	3	3
Male	United States	Student	No	No	No	Go out Every day	Maybe	Yes	No
15552	5622	20746	16091	12102	10627	4523	7661	8138	7701

Gambar 4. Hasil Data Setelah Dibersihkan

3.2.3 Label Encoding

Selanjutnya setelah melakukan pembersihan data, proses selanjutnya adalah melakukan label encoding atau mengubah nilai kategori ke dalam bentuk nilai numerik, dimana berdasarkan Gambar 5 merupakan bentuk tipe awal dari dataset.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8871
Hal 636-647
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

	Gender	Country	Occupation	self_employed	family_history	treatment	Days_Indoors
9	Female	United States	Student	No	No	No	Go out Every day
11	Female	United States	Student	No	No	No	Go out Every day
12	Female	United States	Student	No	No	No	Go out Every day
18	Female	United States	Student	No	No	No	Go out Every day
21	Female	Canada	Student	No	Yes	Yes	Go out Every day

Gambar 5. Sebelum Encoding

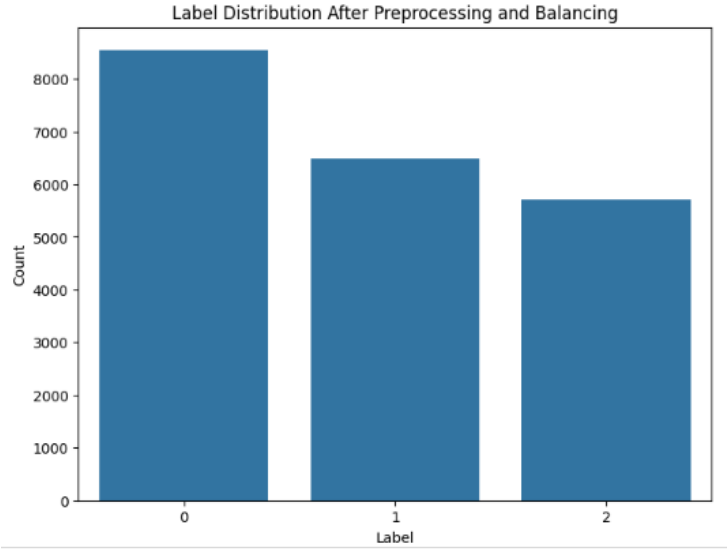
Tujuan dilakukannya label encoding ini adalah untuk mempercepat proses pada saat melakukan *balancing* dengan paramter *class_weight* dan *hyperparameter tuning* menggunakan *GridSearch*, juga mempercepat proses saat *training* pada model LightGBM, hasilnya perbedaannya dapat dilihat pada Gambar 6 berikut.

Gender	Country	Occupation	self_employed	family_history	treatment	Days_Indoors
0	34	0	0	0	1	3
0	34	0	0	1	1	3
0	34	0	0	1	1	3
0	34	0	0	1	1	3
0	34	0	0	1	1	3

Gambar 6. Setelah Encoding

3.2.4 Imbalance Data (class weight dan Hyperparameter Tuning)

Setelah melakukan *encoding*, tahap selanjutnya adalah melakukan *balancing* data menggunakan parameter dari LightGBM yaitu *class_weight* dimana proses tersebut dilakukan bersamaan dengan *hyperparameter tuning* dengan menggunakan teknik *grid search*. Dimana dataset akan diseimbangkan menggunakan parameter class_weight secara otomatis menggunakan "*balanced*", kemudian dilanjut dengan *grid search* untuk mencari parameter terbaik yang akan digunakan dalam proses pelatihan.



Gambar 7. Distribusi Balancing class_weight

Berdasarkan Gambar 7, dengan menggunakan parameter class_weight hasil yang diperoleh adalah data masih tetap sama artinya dengan menggunakan parameter tersebut tidak melakukan menyeimbangkan data. Kemudian, setelah melakukan hyperparameter tuning dengan menggunakan Grid Search diperoleh parameter terbaik seperti pada tabel 5 berikut:

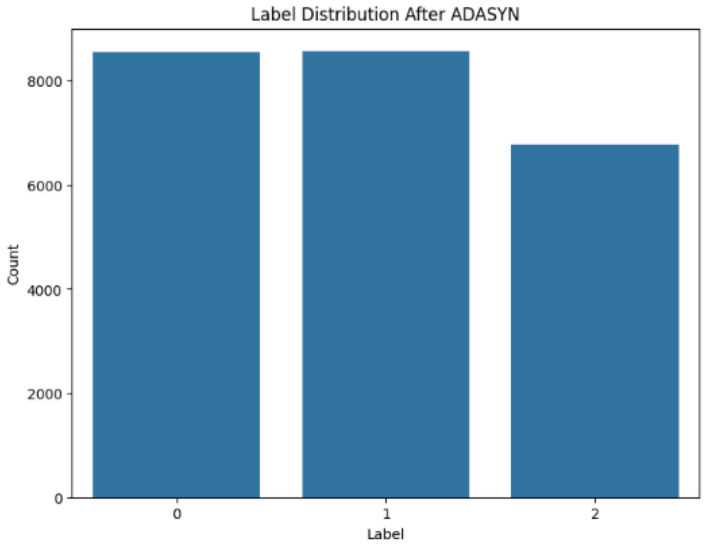
JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8871
Hal 636-647
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tabel 5. Best Parameter

Learning_rate	0.01
Max_depth	20
Min_child_samples	30
N_estimators	100
Num_leaves	70

3.2.5 ADASYN

Berdasarkan pada Gambar 7 dapat diperhatikan jika data tidak mengalami perubahan baik oversampling maupun undersampling. Karena itu data akan dilakukan *balancing* kembali dengan ADASYN, dimana sebelumnya jumlah data pada label 0 = 8551, 1 = 6479 dan 2 = 5716. Setelah melakukan resampling dengan ADASYN jumlah datanya menjadi, 0 = 8551, 1 = 8574, dan 2 = 6782 hasilnya dapat dilihat pada Gambar 8.



Gambar 8. Hasil Balancing ADASYN

Jika diperhatikan pada Gambar 8, label 0 memiliki jumlah data yang tetap ini dikarenakan label tersebut termasuk ke dalam mayaoritas. Sedangkan pada label 1 dan 2 memiliki data yang bertambah namun karena label 1 termasuk ke dalam minoritas, data sintetis pada label tersebut lebih banyak daripada label 2. Kemudian data yang sudah di *resampling* akan dibagi menjadi data latih dan data uji dengan perbandingan 80:20, untuk data training berjumlah 19.125 dan data latih 4.782 data.

3.8 Cross Validation

Tahap selanjutnya adalah melakukan validasi untuk mengevaluasi hasil analisis dengan *Cross Validation*. Tahap ini akan dilakukan tiga percobaan antara lain k=3, k=5 dan k=10, dari 3 percobaan *cross validation* diperoleh hasil rata-rata *F1-weighted score* pada tabel 6 berikut:

Tabel 6. Hasil Cross Validation

K	F1-weighted score
3	0.6804
5	0.6812
10	0.6824

Jika diperhatikan pada tabel 6 diatas nilai rata-rata *F1-weighted score* yang dihasilkan pada *fold* 3, 5 dan 10 mencapai 68%, menunjukkan bahwa model memiliki performa klasifikasi yang cukup stabil terhadap data yang digunakan terutama pada *fold* 10.

3.9 LightGBM

Data latih yang sudah di validasi dengan cross validation akan dilakukan pelatihan dengan menggunakan model LightGBM. Model LightGBM yang digunakan merupakan hasil dari parameter terbaik yang sudah dilakukan pada tahap hyperparameter tuning. Hasil akurasi pelatihan yang diperoleh dengan menggunakan LightGBM dapat dilihat pada tabel 7.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8871
Hal 636-647
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tabel 7. Hasil Training

	Precision	Recall	F1- score	Support
0	0.90	0.54	0.68	6834
1	0.64	0.80	0.71	6823
2	0.61	0.72	0.66	5468
Accuracy			0.69	19125

Berdasarkan pada tabel 7 hasil akurasi pelatihan dengan menggunakan model LightGBM mencapai 0.69. Jika diperhatikan untuk label 0 memiliki precision sebesar 0.90 tetapi hasil recall mencapai 0.54, ini menunjukkan bahwa model yakin dalam memprediksi label 0 namun kurang berhasil dalam mengklasifikasikan data yang termasuk ke dalam label 0 sehingga terdapat data yang tidak berhasil dikenali oleh model. Hal ini terjadi juga pada hasil precision dan recall pada label 1, dimana precision yang diperoleh sebesar 0.64 dan recall mencapai 0.80. Hasil tersebut dapat terjadi karena label 1 sebagai minoritas akan menambah data sintetis atau data buatan, sehingga model lebih terlatih dalam mengenali pola pada label 1 yang menghasilkan recall yang cukup tinggi tetapi hal tersebut menyebabkan model salah dalam mengklasifikasikan sehingga precision yang dihasilkan sebesar 0.64. Hal ini dapat terjadi dari hasil data sintetis yang dihasilkan oleh ADASYN karena tidak cukup merepresentasikan data pada label minoritas.

3.10Evaluasi

Berdasarkan hasil evaluasi pada tabel 8, diperoleh hasil akurasi dengan melakukan balancing dengan *class_weight* dan *hyperparameter tuning* sebesar 0.68. Sedangkan evaluasi pada tabel 9 menunjukkan akurasi yang diperoleh dengan melakukan balancing menggunakan *class_weight, hyperparameter tuning* ADASYN menurun sebesar 0.67. Hasil tersebut dapat dilihat pada tabel 8 dan tabel 9 berikut:

Tabel 8. Akurasi class_weight + hyperparameter tuning

	Precision	Recall	F1- score	Support
0	0.92	0.58	0.71	1711
1	0.54	0.85	0.66	1143
2	0.66	0.66	0.66	1296
Accuracy			0.68	4150

Tabel 9. Akurasi class_weight + hyperparameter tuning + ADASYN

	Precision	Recall	F1- score	Support
0	0.89	0.52	0.65	1717
1	0.64	0.80	0.71	1751
2	0.59	0.72	0.65	1314
Accuracy			0.67	4782

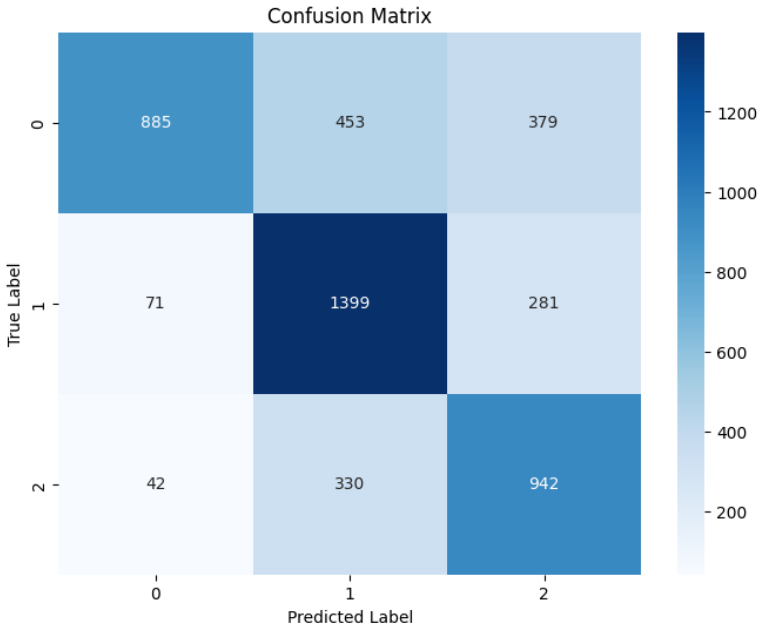
Dari hasil akurasi pada tabel 8 menggunakan parameter class_weight, label yang termasuk minoritas akan ditambahkan bobot kesalahannya lebih besar dari label lain, sehingga model cenderung berhati-hati dalam mengklasifikasikan data minoritas. Namun hal tersebut dapat mempengaruhi precision dan recall dimana jika model terlalu fokus pada label minoritas recall pada label minoritas dapat meningkat tetapi precision yang dihasilkan lebih kecil karena model salah mengklasifikasikan data dari label mayoritas sebagai label minoritas. Sedangkan pada tabel 9 jika menggunakan ADASYN label yang termasuk minoritas akan ditambahkan datanya dengan data sintetis atau data buatan dari ADASYN sehingga model akan lebih mengenali pola label minoritas. Namun hal tersebut dapat menyebabkan model kesulitan membedakan label yang berdekatan sehingga berpengaruh pada hasil precision dan recall. Hasil akurasi dapat menurun karena hasil penambahan data sintetis kurang representatif pada data minoritas dan perbedaan distribusi jika resampling berlebihan dapat menyebabkan adanya overfitting pada data sintetis. Berikut perbandingan pada penelitian sebelumnya dapat dilihat pada tabel 10 berikut:

Tabel 10. Perbandingan Penelitian Terdahulu

Penelitian Saat Ini				**Penelitian Terdahulu [6]**			
Akurasi	**Precision**	**Recall**	**F1-Score**	**Akurasi**	**Precision**	**Recall**	**F1-Score**
0.67	0.71	0.68	0.67	0.88	0.14	1	0.25

Jika dilihat pada tabel 10, terlihat jika hasil precision, recall, dan f1-score cukup seimbang dan mendapatkan hasil rata-rata diatas 0.65, sedangkan untuk penelitian terdahulu memiliki nilai akurasi dan recall yang bagus, tetapi nilai precision dan f1-score yang dihasilkan berada dibawah 30%. Selain itu, terdapat visualisasi confusion matrix pada hasil evaluasi seperti pada Gambar 11 berikut.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 4, Agustus 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i4.8871
Hal 636-647
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

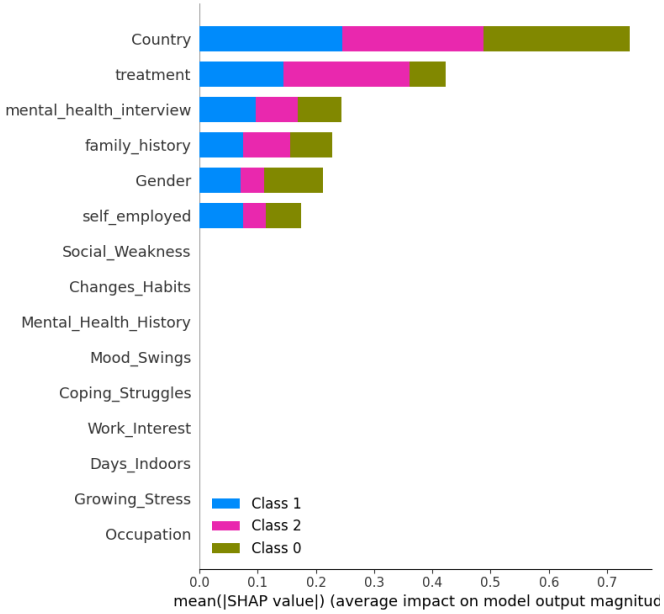


Gambar 9. Confusion Matrix

Dari visualisasi confusion matrix pada Gambar 11 dapat dilihat jika model lebih cenderung mengenali label Not Sure dibandingkan label lainnya. Hal ini dapat terjadi karena kemungkinan karakteristik fitur pada label Not Sure lebih dominan, sehingga menyebabkan adanya ambiguitas dalam proses klasifikasi.

3.11SHAP

Setelah melakukan evaluasi model, selanjutnya adalah menganalisis kontribusi setiap fitur untuk melihat fitur-fitur apa saja yang mempengaruhi hasil klasifikasi pada model LightGBM dengan menggunakan metode SHAP. Metode ini akan menghitung nilai kontribusi pada tiap fitur dengan mempertimbangkan semua kombinasi fitur lainnya. Pada penelitian ini dari 15 fitur yang digunakan dalam klasifikasi terdapat 6 fitur yang memiliki nilai kontribusi tertinggi dapat dilihat pada Gambar 12 berikut.



Gambar 10. Visualisasi SHAP

Berdasarkan gambar diatas diperoleh 6 fitur dengan nilai kontribusi tertinggi antara lain, Country, treatment, mental_health_interview, family_history, Gender, dan self_employed. Dimana fitur Country merupakan fitur paling berkontribusi karena setiap negara memiliki beberapa stigma yang berbeda tentang kesehatan mental. Kemudian ada juga fitur treatment, mental_health_interview dan family_history yang memiliki kontribusi dalam klasifikasi kesehatan mental, karena treatment memiliki pengaruh dalam memberikan keputusan dalam memilih atau tidak dalam melakukan perawatan kesehatan mental. Lalu fitur mental_health_interview menjelaskan terkait pendirian dalam melakukan interview masalah

kesehatan mental dan juga fitur family_history yang menjelaskan bagaimana pengaruh riwayat kesehatan mental dalam lingkungan keluarga untuk meyakinkan pentingnya perawatan kesehatan mental. Terakhir fitur Gender ini membuktikan bahwa kesehatan mental tidak memandang jenis kelamin, baik perempuan atau laki-laki membutuhkan perawatan mengenai kesehatan mental dan fitur self_employed menjelaskan mahasiswa yang memiliki kerja sampingan seperti part time atau freelancer dapat mempengaruhi kesehatan mental, hal ini bisa terjadi karena kurangnya manajemen waktu atau kemampuan individu dalam mengelola stress.

4. KESIMPULAN

Berdasarkan pada penelitian klasifikasi kesehatan mental mahasiswa menggunakan LightGBM dapat disimpulkan bahwa parameter class_weight pada model LightGBM tidak dapat melakukan balancing untuk data yang tidak seimbang, hanya dapat menghitung bobot pada kelas minoritas sehingga model tidak bias terhadap kelas mayoritas selain itu parameter tersebut dapat meningkatkan performa model terhadap kelas yang jarang muncul. Untuk data tidak seimbang dilakukan dengan menggunakan teknik ADASYN yaitu menambahkan data sintetis terhadap data minoritas. Hasil akurasi yang diperoleh dengan menggunakan model LightGBM adalah 0.67, dimana hasil tersebut menunjukkan performa yang baik dalam melakukan klasifikasi kesehatan mental.Saran untuk penelitian selanjutnya adalah melakukan eksplorasi menggunakan model machine learning lainnya seperti XGBOOST, CATBOOST atau deep learning. Selain itu juga eksplorasi metode penyeimbangan data lainnya, seperti SMOTE atau kombinasi lainnya untuk memperoleh hasil yang lebih baik. Kemudian pada penelitian selanjutnya dapat mempertimbangkan klasifikasi dengan menggunakan 6 fitur yang sudah dihitung dengan menggunakan SHAP dan juga dapat menentukan fitur penting dengan menggunakan teknik lainnya selain SHAP.

REFERENCES

[1] Y. R. Shim, R. Eaker, and J. Park, “Mental Health Education, Awareness and Stigma Regarding Mental Illness Among College Students,” J Ment Health Clin Psychol, vol. 6, no. 2, pp. 6–15, Aug. 2022, doi: 10.29245/2578-2959/2022/2.1258.

[2] S. K. Lipson and D. Eisenberg, “Mental health and academic attitudes and expectations in university populations: results from the healthy minds study,” Journal of Mental Health, vol. 27, no. 3, pp. 205–213, May 2018, doi: 10.1080/09638237.2017.1417567.

[3] X. Liu, S. Ping, and W. Gao, “Changes in undergraduate students’ psychological well-being as they experience University Life,” Int J Environ Res Public Health, vol. 16, no. 16, Aug. 2019, doi: 10.3390/ijerph16162864.

[4] M. Gan, S. Pan, Y. Chen, C. Cheng, H. Pan, and X. Zhu, “Application of the machine learning lightgbm model to the prediction of the water levels of the lower columbia river,” J Mar Sci Eng, vol. 9, no. 5, May 2021, doi: 10.3390/jmse9050496.

[5] S. Jain and M. Gangwar, “A Data Mining Analysis Over Psychiatric Database for Mental health Classification,” International Journal on Future Revolution in Computer Science & Communication Engineering, 2018, [Online]. Available: http://www.ijfrcsce.org

[6] B. R. Prasetyo, E. D. Wahyuni, and P. M. Kusumantara, “KOMPARASI PERFORMA MODEL BERBASIS ALGORITMA RANDOM FOREST DAN LIGHTGBM DALAM MELAKUKAN KLASIFIKASI DIABETES MELITUS GESTASIONAL,” Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4817.

[7] L. Qadrini et al., “ENSEMBLE RESAMPLING SUPPORT VECTOR MACHINE, MULTINOMIAL REGRESSION TO MULTICLASS IMBALANCED DATA,” Barekeng, vol. 18, no. 1, pp. 269–280, Mar. 2024, doi: 10.30598/barekengvol18iss1pp0269-0280.

[8] R. Dahlia and C. I. Agustyaningrum, “Perbandingan Gradient Boosting dan Light Gradient Boosting Dalam Melakukan Klasifikasi Rumah Sewa,” Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI), vol. 5, no. 6, pp. 1016–1020, Dec. 2022, doi: 10.32672/jnkti.v5i6.5460.

[9] V. Laijawala, A. Aachaliya, H. Jatta, and V. Pinjarkar, “Classification Algorithms based Mental Health Prediction using Data Mining,” in 2020 5th International Conference on Communication and Electronics Systems (ICCES), IEEE, Jun. 2020, pp. 1174–1178. doi: 10.1109/ICCES48766.2020.9137856.

[10] F. I. Kurniadi and P. D. Larasati, “Light Gradient Boosting Machine untuk Deteksi Penyakit Stroke,” Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan), vol. 6, no. 1, pp. 67–72, Oct. 2022, doi: 10.47970/siskom-kb.v6i1.328.

[11] K. . Aditya Ananta Wisnu Wardana and A. . Mizwar A. Rahim, “Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest untuk Klasifikasi Data Kesehatan Mental,” Jurnal Ilmu Komputer dan Pendidikan, vol. 2, no. Vol. 2 No. 5 (2024): logic : Jurnal Ilmu Komputer dan Pendidikan, pp. 808–818, 2024.

[12] M. Syukron, R. Santoso, and T. Widiharih, “PERBANDINGAN METODE SMOTE RANDOM FOREST DAN SMOTE XGBOOST UNTUK KLASIFIKASI TINGKAT PENYAKIT HEPATITIS C PADA IMBALANCE CLASS DATA,” Jurnal Gaussian, Aug. 2020, doi: https://doi.org/10.14710/j.gauss.9.3.227-236.

[13] Intan Permata and Esther Sorta Mauli Nababan, “Application Of Game Theory In Determining Optimum Marketing Strategy In Marketplace,” JURNAL RISET RUMPUN MATEMATIKA DAN ILMU PENGETAHUAN ALAM, vol. 2, no. 2, pp. 65–71, Jul. 2023, doi: 10.55606/jurrimipa.v2i2.1336.

[14] J. Al Amien, Yoze Rizki, and Mukhlis Ali Rahman Nasution, “Implementasi Adasyn Untuk Imbalance Data Pada Dataset UNSW-NB15 Adasyn Implementation For Data Imbalance on UNSW-NB15 Dataset,” Jurnal CoSciTech (Computer Science and Information Technology), vol. 3, no. 3, pp. 242–248, Dec. 2022, doi: 10.37859/coscitech.v3i3.4339.

[15] D. V. Ramadhanti, R. Santoso, and T. Widiharih, "PERBANDINGAN SMOTE DAN ADASYN PADA DATA IMBALANCE UNTUK KLASIFIKASI RUMAH TANGGA MISKIN DI KABUPATEN TEMANGGUNG DENGAN ALGORITMA K-NEAREST NEIGHBOR," Jurnal Gaussian, vol. 11, no. 4, pp. 499–505, Feb. 2023, doi: 10.14710/j.gauss.11.4.499-505.

[16] W. A. Firmansyach, U. Hayati, and Y. A. Wijaya, "ANALISA TERJADINYA OVERFITTING DAN UNDERFITTING PADA ALGORITMA NAIVE BAYES DAN DECISION TREE DENGAN TEKNIK CROSS VALIDATION," 2023.

[17] T. Chen et al., "Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine," IEEE Access, vol. 7, pp. 150960–150968, 2019, doi: 10.1109/ACCESS.2019.2946980.

[18] M. R. Abbasniya, S. A. Sheikholeslamzadeh, H. Nasiri, and S. Emami, "Classification of Breast Tumours Based on Histopathology Images Using Deep Features and Ensemble of Gradient Boosting Methods," Sep. 2022, doi: 10.48550/arXiv.2209.01380.

[19] E. Ramadanti, D. A. Dinathi, C. Sri, K. Aditya, and R. Chandranegara, "Diabetes Disease Detection Classification Using Light Gradient Boosting (LightGBM) With Hyperparameter Tuning," Jurnal dan Penelitian Teknik Informatika, vol. 8, no. 2, 2024, doi: 10.33395/v8i2.13530.

[20] J. H. Park, H. S. Jo, S. H. Lee, S. W. Oh, and M. G. Na, "A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP," Nuclear Engineering and Technology, vol. 54, no. 4, pp. 1271–1287, Apr. 2022, doi: 10.1016/j.net.2021.10.024.

[21] M. T. Syamkalla, S. Khomsah, and Y. S. R. Nur, "Implementasi Algoritma Catboost Dan Shapley Additive Explanations (SHAP) Dalam Memprediksi Popularitas Game Indie Pada Platform Steam," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 11, no. 4, pp. 777–786, Aug. 2024, doi: 10.25126/jtiik.1148503.