Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 1086-1094
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Klasifikasi Penyakit Gangguan Mental dengan Algoritma LightGBM

Khoirun Nisa
Program Studi Informatika, Universitas Harapan Bangsa, Purwokerto, Indonesia
Email: khoirunnisa@uhb.ac.id

Abstract

As awareness of the importance of mental health increases in society, it unfortunately has not been fully implemented across various industry sectors, including technology. The Open Source Mental Illness (OSMI) survey data measures the level of awareness of mental health among technology industry workers, a group often overlooked. Therefore, it is necessary to develop a classification model to identify patients, which can help doctors initiate early medical treatment. This study aims to evaluate the effectiveness of the LightGBM algorithm model in terms of its accuracy. The dataset used consists of 1259 data points from the OSMI survey. The study found that the best accuracy ratio was achieved with a 90:10 split. The results indicated a training data accuracy of 93%, while the testing data accuracy was 82%.

Keywords*: OSMI, LightGBM, Classification, Accuracy*

Abstrak

Semakin meningkatnya kesadaran akan pentingnya kesehatan mental di tengah masyarakat, sayangnya belum sepenuhnya terimplementasi di berbagai sektor industri, termasuk teknologi. Data Survei Open Source Mental Illness (OSMI) mengukur tingkat kesadaran akan kesehatan mental di kalangan pekerja industri teknologi, yang seringkali diabaikan. Oleh karena itu, perlu dikembangkan model klasifikasi untuk mengidentifikasi pasien yang berguna membantu dokter untuk memulai perawatan medis secara dini. Penelitian ini bertujuan untuk melihat model algoritma LightGBM efektif dalam tingkat akurasinya. Dataset yang digunakan mengambil dari data survei OSMI berjumlah 1259 data. Penelitian ini mengambil rasio data yang menghasilkan akurasi terbaik yaitu 90:10. Hasil penelitian ini didapat bahwa tingkat akurasi data training sebesar 93% namun data testing sebesar 82%.

Kata kunci*: OSMI, LightGBM, Klasifikasi, Akurasi*

1. PENDAHULUAN

Sejalan dengan Rencana Pembangunan Jangka Menengah Nasional (RPJMN) Tahun 2020-2024, Pemerintah Indonesia melalui Kementerian Kesehatan Republik Indonesia (Kemenkes RI) telah menetapkan kesehatan mental sebagai salah satu prioritas utama dalam program kesehatan nasional. Kesehatan mental adalah elemen penting yang mendukung pembangunan negara dan sumber daya manusianya, terutama generasi muda yang akan menentukan arah masa depan Indonesia dan mewujudkan Visi Indonesia Emas 2045. I-NAMHS melaporkan bahwa banyak remaja mengalami masalah kesehatan mental, dengan sekitar satu dari tiga remaja (34,9%) menghadapi masalah tersebut dalam periode 12 bulan [1]. Selain itu, satu dari dua puluh (5,5%) remaja di Indonesia memenuhi kriteria untuk mengalami satu gangguan mental. Berdasarkan data sensus terbaru, prevalensi ini setara dengan 13 juta remaja yang memiliki masalah kesehatan mental dan 2 juta remaja yang memiliki gangguan mental [2].

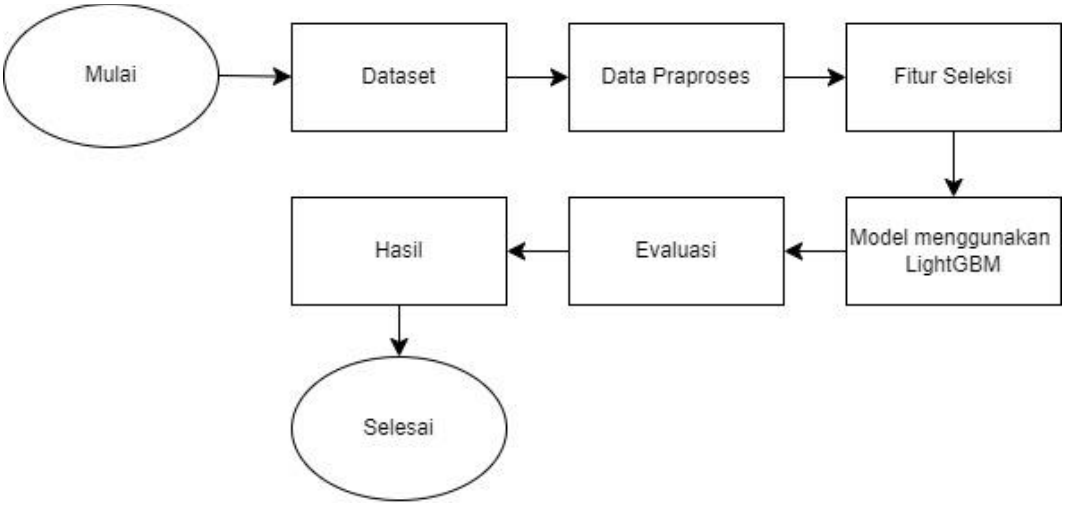
Intervensi dini sangat krusial dalam mengatasi masalah kesehatan mental. Oleh karena itu, perlu mencari solusi yang lebih efektif dalam mendeteksi masalah ini sedini mungkin. Berbagai penelitian telah dilakukan untuk mendukung upaya tersebut. Salah satu, penelitian tentang klasifikasi pasian gangguan jiwa menggunakan algoritma c4.5 memberikan hasil bahwa pada usia 1-6 tahun gangguan jiwa banyak terjadi pada laki-laki, lalu usia 6-8 tahun banyak terjadi pada perempuan, usia 18-45 banyak terjadi pada laki-laki dan usia 45 – 59 tahun pada jenis kelamin perempuan [3]. Penelitian berikutnya menggunakan metode fuzzy madani untuk melakukan pengelompokan gangguan mental pada mahasiswa. Hasil penelitiannya bahwa sebanyak 14 orang menderita skizofrenia paranoid, 12 orang fobia, 28 orang depresi, 16 orang menderita kecemasan, 23 orang OCD, dan 5 orang anti. Proses fuzzy metode mamdani menggunakan 65 aturan dengan variabel output yaitu depresi, kecemasan, skizofrenia, fobia, dan *Obsessive Compultive Disorder* (OCD) [4]. Penelitian sebelumnya terkait klasifikasi gangguan jiwa menggunakan algoritma C5.0 dengan tingkat akurasi sebesar 94,44%. Pengelompokkan sebanyak lima jenis penyakit gangguan jiwa yaitu skizofrenia, depresi, PTSD, bipolar dan psikosis dengan data sejumlah 900 data pasien dibagi menjadi 720 data training dan 180 data testing [5]. Algoritma C4.5 dan naïve bayes digunakan dalam melakukan klasifikasi kesehatan mental karyawan dengan data 1259 orang mengahsilkan nilai gain tertinggi 0,085 [6].

Algoritma KNN dan levenshten distance diterapkan dalam melakukan identifikasi penyakit mental dengan tingkat akurasi sebesar 92,8%. Jenis penyakit mental yang diidentifkasi ada empat yaitu bipolar, anexienty, OCD, dan skizofrenia. Penelitian ini bertujuan untuk mengembangkan sistem yang mampu mengidentifikasi penyakit mental. Sistem ini menggunakan metode TF-IDF untuk memberikan bobot pada kata-kata dari kumpulan keluhan yang diberikan oleh pengguna. Setelah itu, keluhan tersebut akan diklasifikasikan menggunakan algoritma KNN. Selain itu, metode Levenshtein Distance digunakan untuk mengukur jarak antara kata yang diinput oleh pengguna dengan kata-kata yang ada dalam database, serta menghitung jumlah perbedaan antara kedua string dalam bentuk matriks [7]. Penelitian berikutnya menggunakan arsitektur DNN dengan 4 hidden layer, optimasi adagrad, learning rate 0.01 dan epoch 100 menghasilkan tingkat akurasi sebesar 98.25%, presisi 83.00%, recall 98.25% [8]. Penelitian berikutnya menggunakan metode machine learning yaitu KNN, Naïve Bayes, Tree, Random Forest, SVM, AdaBoost dengan hasil tingkat akurasi sebesar 79.6%, 62%, 73%, 72,9%, 70,9% dan 56,2% [9]. Penelitian selanjutnya menggunakan metode random forest dengan hasil tingkat akurasi sebesar 90.83% data yang digunakan sebanyak 120 dengan 17 atribut. Kelompok penyakit mental yaitu depresi, bipolar tipe 1, bipolar tipe 2 dan normal [10]. Melihat pentingnya sistem diagnosis untuk kesehatan mental, banyak penelitian telah dilakukan di bidang ini, mencakup aspek psikologi, kesehatan, dan teknologi. Pada penelitian ini akan mennggunakan metode Light Gradient-Boosting Machine. Metode ini merupakan framework penguat gradien yang menggunakan algoritma *tree based learning*. Framework ini dirancang agar terdistribusi dan efisien dengan keuntungan dalam hal kecepatan pelatihan yang lebih cepat dan efisiensi yang

lebih tinggi, penggunaan memori yang lebih rendah, akurasi yang lebih baik, mendukung pembelajaran paralel, terdistribusi, dan GPU serta ampu menangani data berskala besar [11].

2. METODOLOGI PENELITIAN

Penelitian ini mengacu pada *flowchart* yang ditunjukkan pada Gambar 1, tiap langkah dari flowchart dijelaskan dalam bentuk sub bagian berikut.



Gambar 1. *Flowchart* Alur Penelitian

2.1. Dataset

Tahap awal penelitian ini yaitu pengumpulan dataset yang diambil dari data publik *Kaggle* berasal dari *OpenSource Mental Illness* (OSMI) kumpulan survei yang diperoleh berjumlah 1259 data [12]. Klasifikasi terdiri dari 8 fitur yang digunakan yaitu *Age, Gender, Family History, Benefits, Care_options, Anoymity, Leave,* dan *Work_Interfere.*

2.2. Data Praproses

Tahapan praproses data bertujuan untuk membersihkan data yang akan digunakan dalam analisis. Pembersihan ini melibatkan pengecekan data untuk menemukan data yang null, duplikat, outlier, dan menghapus data yang tidak sesuai dengan kententuan. Lagkah ini dilakukan guna menghindari ketidakakuratan prediksi. Selain itu, proses ini juga mencakup pengurangan fitur dalam dataset yang akan dianalisis, untuk memastikan bahwa data tersebut relevan dengan tujuan analisis[13].

2.3. LightGBM

Light gradient boosting (LightGBM) merupakan salah satu pengembangan dari gradient boosting yang menggunakan algoritma *decision-tree-based learning*. Algoritma LightGBM efisien dalam pelatihan data yang besar dan memiliki performa yang baik dalam hal kecepatan. Algoritma ini juga digunakan dalam menyelesaikan masalah yang berkaitan dengan klasifikasi, regresi, dan klasifikasi.

Model lightGBM diperoleh dengan meminimalkan fungsi boosting loss berdasarkan algoritma penurunan gradien. Setiap model baru ditambahkan, dan *loss function continues* untuk mendapatkan gradien variabel dengan informasi tertinggi. Selain meminimalkan *loss function* dan penerapan *gradient decrease*, lightGBM mempunyai dua fitur utama yaitu metode *leaf-wise tree* dan pengaplikasian algoritma *decision tree* berbasis histogram. Kedua fitur tersebut mampu menangi data berskala besar secara efektif [14].

LightGBM merupakan library machine learning dengan algoritma pembentukan berdasarkan *decision tree gradient-boosting* yang telah dikembangkan untuk memiliki kecepatan yang lebih tinggi. Library LightGBM dapat dipasang di Python atau Google Collaboratory dengan paket pip-python [15]. Persamaan algortima dalam meningkatkan *regresi trees* dapat digeneralkan dalam persamaan (1), dimana model akhir dari model penambahan bertahap sederhana dari nilai b [16].

$$f(x) = \Sigma_b^B =_1 f^b(\mathrm{x}) \quad (1)$$

2.4. Evaluation Metrics

Tahapan evaluasi terhadap metode klasifikasi berguna untuk mengecek kinerja program yang dibuat. Evaluasi menggunakan confusion matrix yang memiliki dua kelas yaitu sehat dan gangguan mental. Berikut ini confusion matrix untuk kelas kesehatan mental tersebut ditunjukkan pada Tabel 1.

Tabel 1. Confusion Matrix Kesehatan Mental

Eksperimen	**Kelas Prediksi**	
Kelas Aktual	**Positif**	**Negatif**
Positif	TP	TN
Negatif	FP	FN

Dimana TP, TN, FN dan FN merupakan singkatan dari *true positives, true negatives, false positives, dan false negatives*. Akurasi, Presisi, *recall,* dan *f1 score* digunakan sebagai metrik pada evaluasi performa dari klasifikasi [17]. Metrik tersebut diformulasikan pada persamaan (2), (3), (4) dan (5).

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FN+FP}\ x\ 100\ \% \quad (2)$$

$$\text{Presisi} = \frac{TP}{TP+FP}\ x\ 100\ \% \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN}\ x\ 100\ \% \quad (4)$$

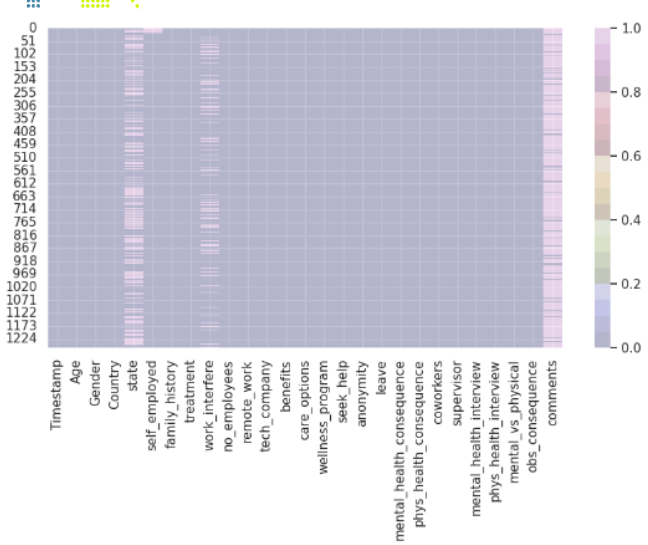
$$\text{F1-Score} = \frac{(1+\beta^2)x\ precision\ x\ recall}{(\beta^2 x\ precision)+\ recall}\ x\ 100\ \% \quad (5)$$

3. HASIL DAN PEMBAHASAN

3.1. *Exploratory Data*

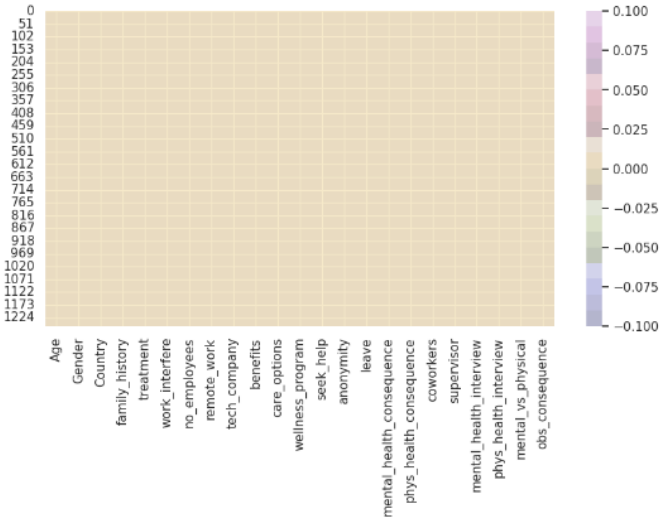
Tahapan Data terdiri dari 1259 responden dengan peubah respon (peubah Y) Yaitu treatment dan 8 peubah prediktor (peubah X). hasil ekplorasi data ditemukan missing value yang cukup tinggi yaitu pada fitur *state* berjumlah 515 (87%) , *self_employed* sejumlah 18 (40.9%), *work_interfere* sebesar 264 (21%) dan *comments* sekitar 1095 (1,4%). *Missing value* divisualkan pada Gambar 2.

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 1086-1094
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik



Gambar 2. *Missing Value*

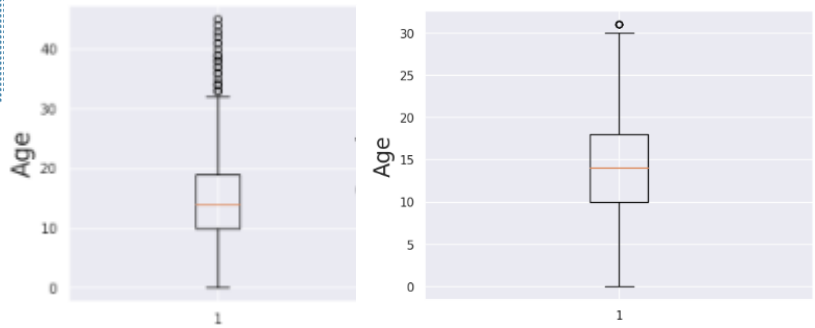
Gambar 3 menunjukan bahwa terdapat banyak data yang dihilangkan pada data karena terjadi *missing value*.



Gambar 3. Setelah pembersihan *Missing Value*

Gambar 4 menunjukan bahwa fitur *Age* disebelah kiri sebelum *outliers* dihilangkan, data masih tersebar dan memiliki *range* yang luas. Setelah *range* pada fitur dihilangkan outlier menurun.

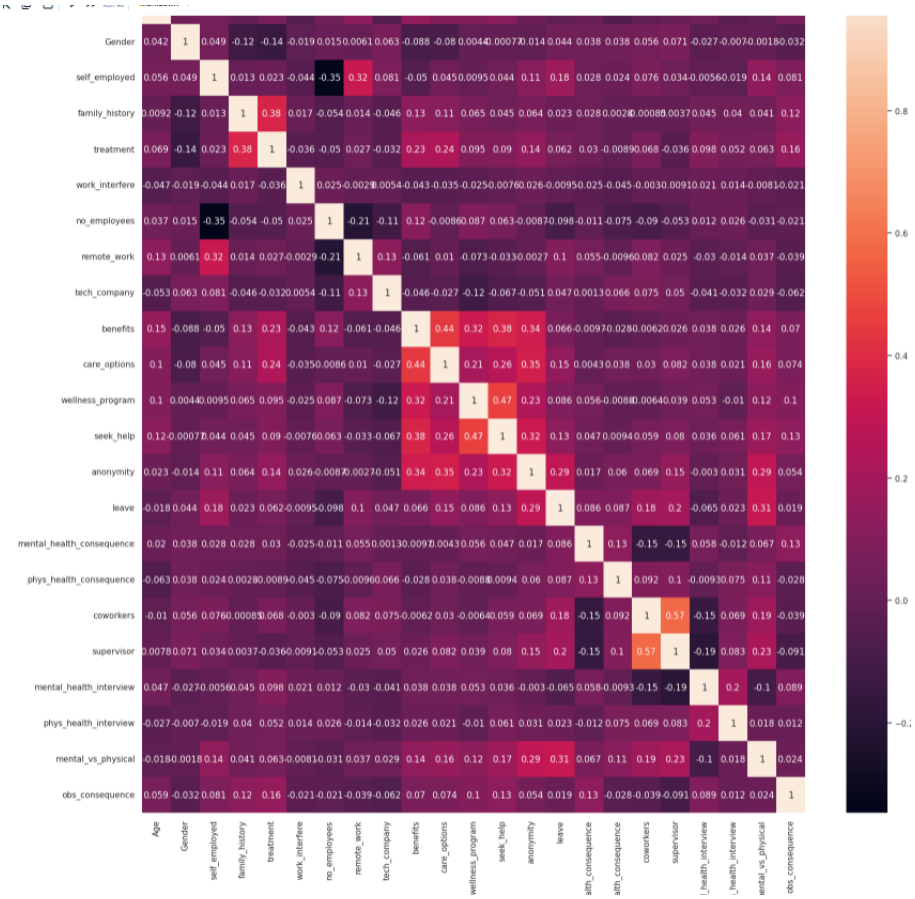
Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 1086-1094
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik



Gambar 4. Box Plot Median Imputation

3.2. Korelasi Variabel

Dataset dipecah menjadi data latih dan data uji, dengan tiga proporsi pembagian yaitu: 70:30, 80:20, dan 90:10, di mana porsi terbesar digunakan untuk data latih. Dataset diperiksa untuk nilai kosong, dan tidak ditemukan data kosong. Analisis hubungan antarvariabel dilakukan menggunakan koefisien korelasi Pearson. Hubungan ini dapat dilihat pada Gambar 5, yang menunjukkan baik relasi langsung maupun tidak langsung antarvariabel. Relasi langsung memiliki rentang nilai antara 0 hingga 1, sedangkan relasi tidak langsung berada dalam rentang 0 hingga -1.



Gambar 5. Korelasi Person antar Variabel

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 2, Agustus 2024, pp 1086-1094
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

3.3. Model Klasifikasi LightGBM

Rasio data training dan data testing yang digunakan dalam penelitian ini membandingkan beberapa rasio untuk menemukan rasio yang paling tepat dapat dilihat pada Tabel 2.

Tabel 2. Split Data Test dan Data Training

Training	Testing	Jumlah data training	Jumlah data testing
70%	30%	881	378
80%	20%	1007	252
90%	10%	1133	126

Data tersebut dilakukan pra-pemrosesan selanjutnya diimplementasikan ke dalam pemodelan algoritma. Pemodelan algoritma yang digunakan adalah algoritma LightGBM. Hasil dari percobaan Pemodelan algoritma lightGBM terhadap split data dijabarkan pada Tabel 3.

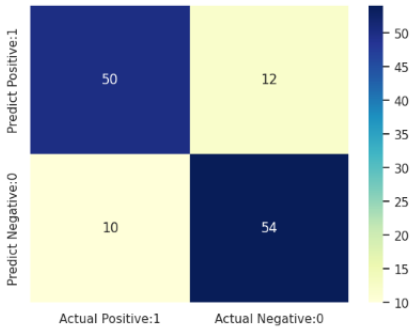
Tabel 3. Hasil Pemodelan Algoritma

Training	Testing	Akurasi Training	Akurasi Testing
70%	30%	94%	79%
80%	20%	94%	77%
90%	10%	93%	83%

Dapat dilihat pada Tabel 3, dari Hasil pemodelan algoritma menggunakan beberapa perbandingan rasio, sehingga didapatkan tingkat akurasi dalam mengklasifikasikan penyakit mental yang paling tinggi adalah klasifikasi dengan menggunakan perbandingan rasio 90% untuk data training dan 10% untuk data testing. Oleh karena itu, dalam penelitian ini menggunakan perbandingan rasio tersebut.

3.4. Evaluasi

Implementasi algoritma LightGBM menghasilkan jumlah data True Negative (TN) sebanyak 54, False Negative (FN) sebanyak 10, True Positive (TP) sebanyak 50, False Positive (FP) sebanyak 12 data. Detail dari *Confusion Matrix* dapat dilihat pada Gambar 6.



Gambar 6. *Confusion Matrix*

Setelah mendapatkan nilai dari masing-masing TP, TN, FP, dan FN, selanjutnya dapat menentukan nilai dari precision, recall, f1-Score, dan akurasi pada Gambar 7.

	precision	recall	f1-score	support
0	0.83	0.81	0.82	62
1	0.82	0.84	0.83	64
accuracy			0.83	126
macro avg	0.83	0.83	0.83	126
weighted avg	0.83	0.83	0.83	126

Gambar 7. Hasil pengujian algoritma

4. SIMPULAN

Hasil penelitian yang dilakukan, disimpulkan bahwa model klasifikasi yang dihasilkan dari prediksi data survei kesehatan pada pekerja di industri teknologi dapat digunakan untuk menentukan apakah responden di industri tersebut sudah sadar atau belum terhadap pengaruh kesehatan mental. Selain itu, analisis klasifikasi menggunakan algoritma LightGBM menghasilkan akurasi 83% dengan pembagian data 90% untuk pelatihan dan 10% untuk pengujian, yang merupakan akurasi tertinggi yang dicapai setelah mencoba beberapa pembagian data lainnya.

DAFTAR PUSTAKA

[1] Indonesia National Adolescent Mental Health Survei, "National Adolescent Mental Health Survey (I-Namhs) Laporan Penelitian," *Ment. Health (Lond).*, P. Xviii, 2022, [Online]. Available: Https://Qcmhr.Org/Outputs/Reports/12-I-Namhs-Report-Bahasa-Indonesia.

[2] Statistics Indonesia, "Hasil Sensus Penduduk 2020," *Statistics Indonesia*, 2021. .

[3] H. Hendra, M. Muhaemin, And S. Santosa, "Klasifikasi Pasien Gangguan Jiwa Menggunakan Algoritma C4. 5 Sebagai Dasar Pengambilan Keputusan Kesehatan Jiwa," *Pros. Semin. Nas. …*, 2023, [Online]. Available: Https://Jurnal.Umj.Ac.Id/Index.Php/Semnaslit/Article/View/19400%0ahttps://Jurnal.Umj.Ac.Id/Index.Php/Semnaslit/Article/Download/19400/9465.

[4] T. Tanugeraha, A. J. Santoso, And S. P. Adithama, "Pengelompokan Gangguan Kesehatan Mental Mahasiswa Yang Sedang Menempuh Skripsi Dengan Metode Fuzzy Mamdani," *J. Inform. Atma Jogja*, Vol. 4, No. 1, Pp. 77–84, 2023, Doi: 10.24002/Jiaj.V4i1.7445.

[5] N. Anthira And Suendri, "Penerapan Data Mining Pada Klasifikasi Gangguan Jiwa Menggunakan Algoritma C5.0 Di Rsj. Mahoni Kota Medan," *J. Tek.*, Vol. 18, No. X, Pp. 571–582, 2024.

[6] M. Nonsi Tentua, V. Fidiantoro, And P. Feri Ariyanto, "Metode C4.5 Dan Naive Bayes Untuk Klasifikasi Kesehatan Mental Karyawan Di Tempat Kerja," *J. Din. Inform.*, Vol. 11, No. 2, Pp. 1–16, 2022, [Online]. Available: Https://Www.Kaggle.Com/Osmi/Mental-.

[7] Y. Rahma, A. Prasetiadi, And M. Wibowo, "Identification Of Mental Ilness From Patient Diseases Using Knn And Levenshtein Distance Algorithm," *J. Tek. Inform.*, Vol. 3, No. 5, Pp. 1363–1372, 2022, Doi: 10.20884/1.Jutif.2022.3.5.371.

[8] M. A. Azis, A. Fauzi, G. Ginabila, And I. Nawawi, "Klasifikasi Human Stress Menggunakan Adagrad Optimization Untuk Arsitektur Deep Neural Network," *J.*

Tek. Inform. Unika St. Thomas, Vol. 07, Pp. 56–62, 2022, Doi: 10.54367/Jtiust.V7i1.1916.

[9] M. Anastasia, V. S. Maulivia, And S. Suharjito, "Metode Pembelajaran Mesin Untuk Menilai Data Depresi Dan Kesehatan Mental," *Intecoms J. Inf. Technol. Comput. Sci.*, Vol. 7, No. 3, Pp. 606–612, 2024, Doi: 10.31539/Intecoms.V7i3.9584.

[10] A. Priyono;, M. Shodiq, D. P. Alvinsyah, And S. A. Hidayah, "Metode Random Forest Untuk Memudahkan Klasifikasi Diagnosis Penyakit Mental," *J. Inform. Medis*, Vol. 2, No. 1, Pp. 10–13, 2024.

[11] "Lightgbm." Https://Lightgbm.Readthedocs.Io/En/Stable/ (Accessed Jul. 29, 2024).

[12] Osmi, "Mental-Health-In-Tech-Survey." Https://Www.Kaggle.Com/Datasets/Osmi/Mental-Health-In-Tech-Survey (Accessed Jul. 29, 2024).

[13] A. P. S. Iskandar *Et Al.*, *Teknologi Big Data (Pengantar Dan Penerapan Teknologi Big Data Di Berbagai Bidang)*, Pertama. Yogyakarta: Pt. Green Pustaka Indonesia, 2024.

[14] H. Zeng *Et Al.*, "A Lightgbm-Based Eeg Analysis Method For Driver Mental States Classification," Vol. 2019, 2019, Doi: 10.1155/2019/3761203.

[15] R. Latifah, G. Erda, And S. S. Program, "Application Of The Lightgbm Algorithm In The Classification," Vol. 4, No. 1, Pp. 9–15, 2024.

[16] I. Wardhana, M. Ariawijaya, V. A. Isnaini, And R. P. Wirman, "Gradient Boosting Machine, Random Forest Dan Light Gbm Untuk Klasifikasi Kacang Kering," *Resti*, Vol. 6, Pp. 92–99, 2022.

[17] N. Istiana And A. Mustafiril, "Perbandingan Metode Klasifikasi Pada Data Dengan Imbalance Class Dan Missing Value," *Urnal Inform.*, Vol. 10, No. 2, Pp. 101–108, 2023, Doi: Https://Doi.Org/10.31294/Inf.V10i2.15540.