TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi
Vol. 4 No. 2 (Desember 2024) – ***Edisi Khusus SEMNASTIK 2024***
https://doi.org/10.46880/tamika.Vol4No2(SEMNASTIK).pp40-48 e-ISSN: 2828-1276

PREDIKSI GANGGUAN KESEHATAN MENTAL PADA KALANGAN MAHASISWA MENGGUNAKAN METODE *PSEUDO-LABELING* DAN ALGORITMA REGRESI LOGISTIK

Anggraini Puspita Sari✉, Dwi Arman Prasetya, Firza Prima Aditiawan, Muhammad Muharrom Al Haromainy
Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia
Email: anggraini.puspita.if@upnjatim.ac.id

ABSTRACT

Mental illness is a health condition that alters a person's thoughts, feelings, or behaviors, leading to distress and difficulty in maintaining a normal life. Mental health issues should not be taken lightly due to the challenges associated with diagnosis. Many students tend to experience mental health problems at various stages of their education, from diploma programs to doctoral studies. This situation becomes more critical as students approach the end of their studies and anticipate future prospects. This article explores the mental health status of students through symptoms, using logistic regression methods for prediction based on the dataset used. In this study, two types of data are employed: labeled dataset and unlabeled dataset, which are combined to create a semi-supervised learning approach. Labeled dataset is classified using a logistic regression algorithm, while unlabeled dataset employs the pseudo-labeling method. The analysis and modeling of the dataset indicate that the comparison between labeled and unlabeled dataset can significantly affect accuracy and processing time. Furthermore, the use of the pseudo-labeling method with the logistic regression algorithm is well-suited for the mental health case study, achieving an accuracy of 98% with a labeled to unlabeled dataset ratio of 1:2.
Keyword: Mental Health, Pseudo-Labeling, Logistic Regression.

ABSTRAK

Penyakit mental merupakan kondisi kesehatan yang mengubah pemikiran, perasaan, atau perilaku seseorang sehingga orang tersebut tertekan dan kesulitan untuk menjalani kehidupan normal. Permasalahan kesehatan mental tidak boleh dianggap remeh dikarenakan sulit didiagnosis. Banyak mahasiswa cenderung mengalami masalah kesehatan mental pada berbagai tahap pendidikan baik dari jenjang pendidikan d3 hingga S3. Ini menjadi lebih kritis saat mahasiswa mendekati akhir studi dan menantikan prospek masa depan. Artikel ini mengeksplorasi keadaan kesehatan mental mahasiswa melalui gejala dengan menggunakan metode regresi logistik untuk proses prediksi berdasarkan dataset yang digunakan. Pada penelitian ini digunakan dua jenis dataset, yakni dataset labeling dan dataset unlabeling dimana kedua jenis dataset tersebut digabungkan sehingga menjadi Semi-Supervised Learning. Pada dataset labeling digunakan klasifikasi dengan algoritma Regresi Logistik sedangkan pada dataset unlabeling digunakan metode Pseudo Labeling. Dari hasil proses analisis dan pemodelan dataset yang telah dilakukan terlihat bahwa penggunaan perbandingan antara dataset labeling dan unlabeling dapat mempengaruhi hasil akurasi serta lama waktu pemrosesan. Selain itu, penggunaan metode pseudo-labeling dengan algoritma regresi logistik sangat cocok untuk studi kasus kesehatan mental yang menghasilkan akurasi mencapai 98% dengan perbandingan proporsi dataset labeling dan unlabeling sebesar 1:2.
Kata Kunci: Kesehatan Mental, Pseudo-Labeling, Regresi Logistik.

PENDAHULUAN

Penyakit mental merupakan kondisi kesehatan yang mengubah pemikiran, perasaan, atau perilaku seseorang sehingga orang tersebut tertekan dan kesulitan untuk menjalani kehidupan normal (Ridlo, 2020; Aloysius & Salvia, 2021). Permasalahan kesehatan mental tidak boleh dianggap remeh dikarenakan sulit didiagnosis. Banyak mahasiswa cenderung mengalami masalah kesehatan mental pada berbagai tahap pendidikan baik dari jenjang pendidikan d3 hingga S3. Ini menjadi lebih kritis saat mahasiswa mendekati akhir studi dan menantikan prospek masa depan.

WHO menyatakan bahwa satu dari empat orang di dunia akan dipengaruhi oleh gangguan mental atau neurologis di beberapa titik dalam kehidupan mereka. Sekitar 450 juta orang saat ini menderita kondisi seperti itu, menempatkan gangguan mental di antara penyebab utama kesehatan buruk dan cacat di seluruh dunia. Salah satu contoh gangguan mental adalah depresi.

WHO menunjukkan bahwa depresi adalah penyakit umum di seluruh dunia, dengan lebih dari 300 juta orang terkena dampaknya (Aloysius & Salvia, 2021; Kesehatan, 2018; Florensa, Hidayah, Sari, Yousrihatin, & Litaqia, 2023).

Di Indonesia, *pravelensi* depresi pada penduduk umur >15 tahun adalah 6,1 persen. Selain itu, *prevalensi* rumah tangga yang mempunyai anggota rumah tangga yang mengalami gangguan jiwa *Skizofrenia*/Psikosis adalah sebesar 6,7 persen . Berdasarkan data Riskesdas, dari tahun 2013 ke tahun 2018 pravelensi GME (Gangguan Mental Emosional) kelompok umur 15-24 tahun mengalami peningkatan yang paling signifikan dibanding kelompok umur lainnya (Florensa, Hidayah, Sari, Yousrihatin, & Litaqia, 2023; Dzil Kamalah & Nafiah, 2023).

Rentang umur mahasiswa termasuk ke dalam kelompok umur 15-24 tahun. Mahasiswa berada pada tahap remaja akhir (*adolescence*: 10-20 *years*) dan dewasa awal (*early adulthood* : *20's and 30's*). Rentang usia mahasiswa berada pada batasan remaja akhir dan dewasa awal, dimana masa ini merupakan masa kondisi mental yang tidak stabil, diiringi dengan konflik dan tuntutan serta perubahan suasana hati. Apabila individu yang mengalami masa tersebut tidak dapat mengontrol hal-hal yang terjadi, maka dapat menimbulkan masalah kesehatan mental yang akan memempengaruhi kesehatannya secara keseluruhan (Kesehatan, 2018; Endriyani, Lestari, Lestari, & Napitu, 2022; Putri, Wibhawa, & Gutama, 2020).

Penelitian yang dilakukan oleh WHO dalam WHO *World Mental Health International College Student project* yang meneliti sembilan belas universitas di delapan negara ditemukan bahwa 35 persen mahasiswa seumur hidupnya mengalami setidaknya satu mental disorder DSM-IV yaitu *anxiety*, *mood*, atau *substance disorder* dimana dan 31,4 persen mengalaminya dalam rentang 12 bulan terakhir. Penelitian juga pernah dilakukan mengenai masalah kesehatan jiwa mahasiswa baru di sebuah universitas di Jakarta. Hasil dari penelitiannya menunjukkan bahwa 12,69 persen mahasiswa mengalami masalah kejiwaan (Florensa, Hidayah, Sari, Yousrihatin, & Litaqia, 2023; Nurhaeni, Marisa, & Oktiany, 2022; Pratiwi & Rusinani, 2022).

Gangguan kesehatan mental berdampak kepada berbagai aspek kehidupan manusia. Ketika depresi bertahan lama dan dengan intensitas sedang atau berat, depresi dapat menjadi kondisi kesehatan yang serius. Hal ini dapat menyebabkan orang yang terkena sangat menderita dan tidak dapat berfungsi dengan baik di tempat kerja, di sekolah dan di keluarga. Hal terburuknya depresi dapat menyebabkan bunuh diri. Hampir 800.000 orang meninggal karena bunuh diri setiap tahun. Bunuh diri menempati urutan kedua penyebab utama kematian pada usia 15-29 tahun (Dzil Kamalah & Nafiah, 2023; Radiani, 2019; Rahmawaty, Silalahiv, & Mansyah, 2022).

Seiring dengan perkembangan ilmu pengetahuan dan teknologi informasi, gangguan kesehatan mental ini bisa dilakukan deteksi dini dengan menerapkan metode pada *machine learning* atau *artificial intelligence* (AI). Metode yang cocok untuk deteksi dini tersebut dapat menerapkan metode yang menggunakan klasifikasi, antara lain, *Naïve Bayes*, *Random Forest*, *Support Vector Machine*, *Decision Tree*, *K-Nearest Neighbor*, regresi logistik dan sebagainya.

Regresi logistik memiliki beberapa kelebihan, yaitu (a) *Simplicity and Interpretability*: mudah dipahami dan hasilnya dapat diinterpretasikan secara jelas, (b)*Handling Binary Outcomes*: sangat efektif untuk memodelkan variabel dependen biner, yang umum digunakan dalam penelitian kesehatan mental, (c) *Efficiency with Large Datasets*: mampu menangani dataset besar dan kompleks dengan cepat, menjadikannya pilihan yang baik untuk analisis data di bidang kesehatan, (d)*Robustness*:cenderung tahan te rhadap outlier, yang dapat meningkatkan keandalan model dalam analisis kesehatan mental, (e) *Flexibility*: dapat dengan mudah diperluas untuk menangani masalah yang lebih kompleks, seperti regresi logistik multinomial atau model dengan interaksi antara variabel (Alwi, Ermawati, & Husain, 2018; Pratama, Nurcahyo, & Firgia, 2023; Novitasari & dkk, 2019). Berdasarkan kelebihan regresi logistik tersebut akan diterapkan untuk memprediksi gangguan kesehatan mental mahasiswa. Hal ini diharapkan dapat menghasilkan akurasi yang tinggi sehingga deteksi dini pada gangguan kesehatan mental dapat terdeteksi sesuai yang diinginkan. Selain itu, penelitian ini menggunakan dua jenis data, yakni data labeling dan data unlabeling dimana kedua jenis data tersebut digabungkan sehingga menjadi Semi-Supervised Learning. Pada data labeling digunakan klasifikasi dengan algoritma Regresi Logistik sedangkan pada data unlabeling digunakan metode pseudo-labeling sehingga diharapkan hasil prediksi dengan perpaduan antara data labeling dan unlabeling dapat meningkatkan hasil akurasi secara signifikan.

METODE PENELITIAN

Alur penelitian pada artikel ini ditunjukkan dalam Gambar 1. Alur penelitian ini dimulai dengan data collection, *Exploratory Data Analysis* (EDA), pre-

processing, Feature Engineering, pemodelan dan diakhiri dengan evaluasi model.

Data Collection

Pengumpulan data merupakan proses mengumpulkan informasi atau fakta yang relevan dengan tujuan penelitian atau analisis tertentu. Pengumpulan data merupakan langkah yang krusial dalam proses penelitian karena data yang tidak akurat dan tidak representatif menghasilkan temuan yang tidak valid (Sari, et al., 2022; Juwairi, 2019). Dataset yang digunakan mengenai "*Mental Health*" dan menggunakan dua jenis, yaitu dataset *labeling* yang bersumber dari "*Kaggle*" dan dataset *unlabeling* yang didapatkan dari menyebarkan kuesioner pada lingkungan mahasiswa Universitas Pembangunan Nasional "Veteran" Jawa Timur (UPNVJT). Kuesioner yang disebarkan berisi variabel-variabel yang sama dengan variabel pada dataset yang didapatkan dari "*Kaggle*". Data hasil kuesioner tersebut nantinya akan digabungkan dengan dataset yang lama dengan metode *Pseudo-Labeling*. Dataset yang digunakan merupakan dataset mengenai variabel-variabel tentang sebab akibat kesehatan mental dan diagnosis pada kesehatan mental. Dataset *unlabeling* berjumlah 300 dan dataset *labeling* berjumlah 637.

EDA

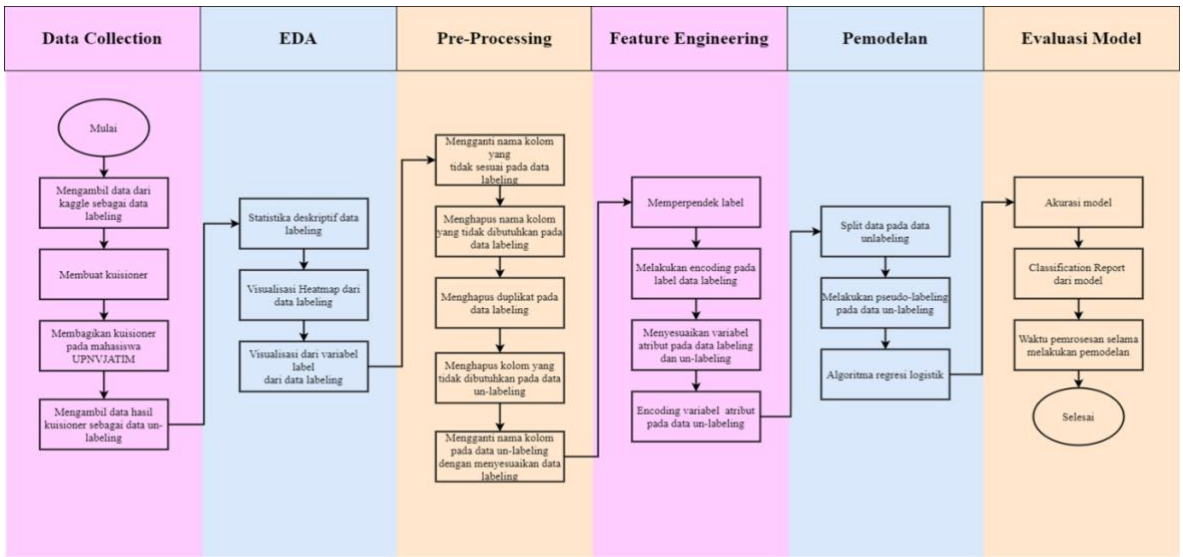
EDA merupakan salah satu proses analisis data yang bertujuan untuk mengungkap dan memahami karakteristik, pola, dan hubungan dalam dataset secara visual dan deskriptif. Tujuan dari EDA adalah untuk mendapatkan pemahaman yang mendalam tentang data yang ada, mengidentifikasi pola menarik, dan menghasilkan wawasan yang dapat membantu dalam pengambilan keputusan atau pengembangan model lebih lanjut (Juwairi K. P., 2019).

Statistika Deskriptif

Statistika Deskriptif digunakan untuk menggambar dan menganalisis data secara numerik untuk menyajikan informasi secara ringkas dan mudah dipahami, serta memberikan gambaran mengenai data yang ada dalam bentuk sampel maupun populasi. Dengan menggunakan Statistika Deskriptif, maka dapat memahami rata-rata, variasi, dan distribusi data yang memungkinkan untuk membuat kesimpulan, dan mengambil keputusan kedepannya.

Visualisasi Data

Visualisasi data digunakan untuk menyajikan data dengan lebih detail untuk mengidentifikasi pola, tren, anomali atau hubungan antara variabel secara lebih jelas. Visualisasi juga dapat mempermudah dalam memahami informasi melalui gambar dibandingkan dengan menggunakan tabel atau angka. Pada Penelitian ini menggunakan visualisasi *Heatmap* untuk memvisualisasikan dari dataset *labeling* dengan menggunakan warna untuk mewakili nilai-nilai yang menunjukkan pola atau korelasi dalam data yang kompleks. Selain itu, penelitian ini juga menggunakan visualisasi dari variabel label dari dataset *labeling* untuk membantu dalam evaluasi kinerja model.



Gambar 1. Alur Penelitian

Pre-Processing

Pre-Processing merupakan proses untuk menyiapkan data mentah untuk menjadi data siap pakai yang didalamnya terdapat beberapa serangkaian yang dilakukan sesuai kebutuhan sebelum data tersebut digunakan untuk proses analisis.

- Menghapus data duplikat
  Memastikan tidak terdapat data duplikat digunakan untuk mendapatkan wawasan yang jelas tentang data yang sedang dianalisis. Apabila terdapat data yang duplikat, nantinya akan mempengaruhi validitas dan hasil dari pemodelan atau analisis yang dilakukan yang akan memberikan hasil yang tidak akurat. Data yang terduplikat mempengaruhi interpretasi hasil.
- Mengganti nama kolom yang tidak dibutuhkan
  Mengganti nama kolom yang tidak dibutuhkan pada data *unlabeling* digunakan untuk menyesuaikan data *labeling*. Dengan mengganti nama kolom pada data *unlabeling* nantinya akan mempermudah proses integrasi atau penggabungan data *labeling* dan *unlabeling* untuk analisis lebih lanjut maupun untuk persiapan pemodelan agar memperoleh hasil yang lebih akurat dan pengambilan keputusan yang lebih baik secara keseluruhan.
- Menghapus nama kolom yang tidak dibutuhkan
  Menghapus nama kolom yang tidak dibutuhkan pada data *labeling* maupun *unlabeling* dapat mempercepat proses analisis dan pemodelan, memperkecil dimensi data untuk membantu analisis data yang efisien, pemodelan yang akurat dan menghasilkan data yang lebih baik.

Feature Engineering

Feature Engineering merupakan salah satu tahap pada *machine learning* yang didalamnya dilakukan proses ekstraksi fitur dari data. Tahapan ini juga berguna untuk meningkatkan akurasi model. Dalam tahapan ini pula, akan dipilih variabel prediktor yang paling berpengaruh terhadap model (Rajoub, 2020; Sari, et al., 2022). Pada penelitian ini, ada beberapa tahap *feature engineering* yang akan dilakukan. Berikut adalah tahapan *feature engineering* yang dilakukan pada pengolahan dataset *mental health*.

- Memperpendek Label
  Nama kolom atau label adalah atribut atau identitas di setiap kolomnya. Hal ini yang membedakan kolom satu dengan lainnya. Oleh sebab itu, pemberian label sangat berpengaruh pada proses analisis. Hal ini dikarenakan, semakin panjang label akan mempersulit proses analisis. Pada penelitian ini, diperlukan adanya tahap untuk memperpendek label. Hal ini bertujuan untuk menyingkat beberapa label yang ada pada setiap kolom agar memudahkan proses analisis.
- *Encoding Data*
  Encoding data adalah tahapan untuk mengubah data menjadi suatu kode-kode tertentu. Ini biasanya diperlukan ketika data bertipe kategorik. Hal ini bertujuan untuk memudahkan proses analisis. Pada penelitian ini, diperlukan adanya encoding data. Hal ini disebabkan data yang digunakan bertipe kategorik yaitu berupa "Ya/Yes" dan "No/Tidak". Oleh sebab itu, data perlu ditransformasikan menjadi "1" dan "0" untuk memudahkan proses analisis.
- Penyesuaian Variabel
  Penyesiuan variabel adalah tahap dimana peneliti harus mengecek kesamaan label antara dua dataset atau lebih yang digunakan. Hal ini bertujuan untuk memudahkan proses analisis. Pada penelitian ini, terdapat 2 dataset yaitu labeling dari Kaggle dan unlabeling dari kuesioner. Keduanya tentu mempunyai kolom-kolom yang berbeda. Oleh sebab itu, perlu adanya penyesuaian variabel menjadi label dataset 1 atau label dataset 2.

Pemodelan

Pemodelan adalah tahap untuk membangun model dari suatu algoritma sehingga dapat diketahui prediksinya. Ada beberapa tahapan yang dilakukan pada pemodelan data. Berikut adalah tahapan yang dilakukan dalam pemodelan data.

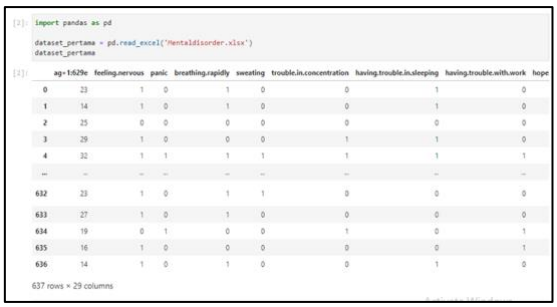
- Split data
  Splitting data adalah tahapan dimana membagi data menjadi 2 kelompok yaitu data *training* dan data *testing*. Data *training* adalah data yang digunakan untuk melatih *machine* sehingga *machine* dapat belajar mengenali karakteristik dari data. Sedangkan data *testing* adalah data yang digunakan untuk menguji *machine* sehingga dapat diketahui *machine* ini belajar dan seakurat apa *machine* dapat memprediksi.
- *Pseudo-labeling*
  Pesudo-labeling (juga dikenal sebagai *Self-Training*) adalah teknik dalam pembelajaran mesin *semi supervised* yang melibatkan penggunaan data yang tidak diberi label (*unlabeled data*) untuk melatih model dengan memberikan label *pseudo* (*pseudo-label*) pada data tersebut. *Pseudo-Labeling* digunakan untuk memanfaatkan data yang tidak diberi label untuk meningkatkan kinera model pembelajaran mesin dengan memberikan label *pseudo* pada data tersebut. Dengan menggunakan *Psedo-Labeling* juga akan membantu dalam meningkatkan kinerja model seperti Regresi Logistik dengan memanfaatkan informasi yang terkandung dalam data yang tidak diberi label.
- Algoritma Regresi Logistik
  Algoritma regresi logistik adalah algoritma yang merupakan implementasi dari regresi logistik.

Evaluasi Model

Evaluasi model adalah sebuah kegiatan untuk memilih model mana yang paling cocok digunakan pada dataset. Ada beberapa nilai yang perlu diperhatikan dalam pemilihan model, seperti akurasi, presisi, recall, dan *F1-score*. Pada penelitian ini, dengan memanfaatkan fitur *classification_report* yang ada pada *library sklearn* sehingga dapat diketahui nilai akurasi, presisi, *recall*, dan *F1-score*.

HASIL DAN PEMBAHASAN

Pada penelitian ini menggunakan dataset *semi-supervised*. Dataset *semi-supervised* merupakan dataset gabungan antara dataset *labelling* dan dataset *unlabelling*. Dataset *labeling* diambil dari situs data yaitu "Kaggle". Dataset tersebut memiliki 637 *record* data seperti yang ditunjukkan Gambar 2. Sedangkan, untuk dataset *unlabelling* diambil dari survei kepada mahasiswa di lingkungan UPNVJT yang disebar melalui kuisioner. Berdasarkan hasil penyebaran kuesioner tersebut diperoleh 300 *record* data seperti pada Gambar 3.



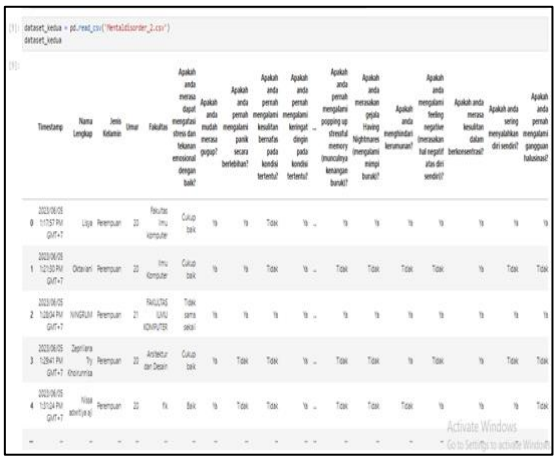
```
import pandas as pd

dataset_pertama = pd.read_excel('Mentaldisorder.xlsx')
dataset_pertama
```

	ag=1:629e	feeling.nervous	panic	breathing.rapidly	sweating	trouble.in.concentration	having.trouble.in.sleeping	having.trouble.with.work	hope
0	23	1	0	1	0	0	1	0	
1	14	1	0	1	0	0	1	0	
2	25	0	0	0	0	0	0	0	
3	29	1	0	0	0	1	1	0	
4	32	1	1	1	1	1	1	1	
...	...	...	...	...	...	...	...	...	
632	23	1	0	1	1	0	0	0	
633	27	1	0	1	0	0	0	0	
634	19	0	1	0	0	1	0	1	
635	16	1	0	0	0	0	0	1	
636	14	1	0	1	0	0	1	0	

637 rows × 29 columns

Gambar 2. Dataset Labeling

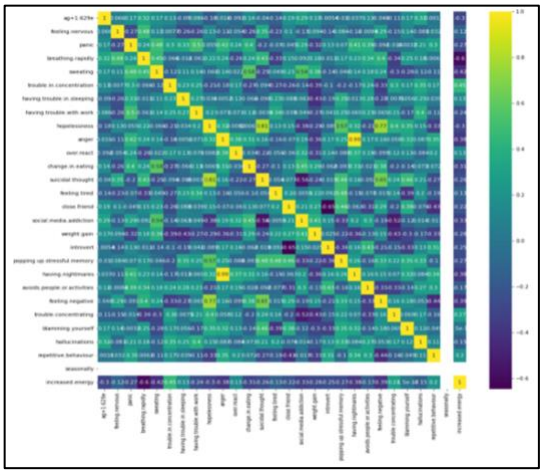


```
dataset_lesua = pd.read_csv('Mentaldisorder_2.csv')
dataset_lesua
```

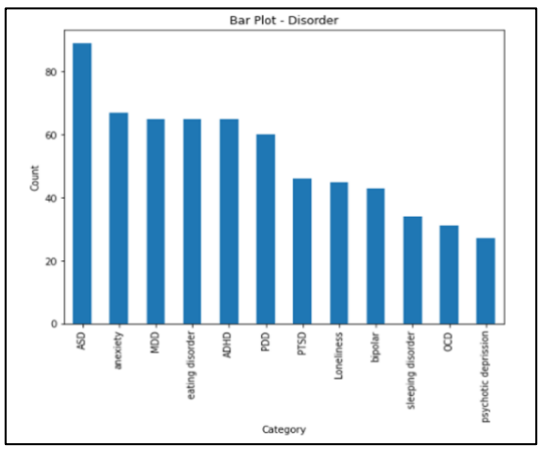
	Timestamp	Nama Lengkap	Jenis Kelamin	Umur	Fakultas	Apakah anda merasa dapat mengatasi stres dan tekanan emosional dengan baik?	Apakah anda mudah merasa gugup?	Apakah anda pernah mengalami panik secara berlebihan?	Apakah anda pernah mengalami kesulitan bersantai pada kondisi tertentu?	Apakah anda pernah mengalami keringat ... dingin pada kondisi tertentu?	Apakah anda pernah mengalami popping up stressful memory (munculnya kenangan buruk)?	Apakah anda merasakan gejala Having Nightmares (mengalami mimpi buruk)?	Apakah anda menghindari keramaian?	Apakah anda mengalami feeling negative (merasakan hal negatif atau diri sendiri)?	Apakah anda merasa kesulitan dalam berkonsentrasi?	Apakah anda sering menyalahkan diri sendiri?	Apakah anda pernah mengalami gangguan halusinasi?
0	2023/06/05 11:57:51 PM GMT+7	Liga	Perempuan	22	Fasilkom: Ilmu komputer	Cukup baik	Ya	Ya	Tidak	Ya	...	Ya	Ya	Ya	Ya	Ya	Ya	Ya
1	2023/06/05 12:10:10 PM GMT+7	Octavian	Perempuan	22	Ilmu Komputer	Cukup baik	Ya	Ya	Tidak	Ya	...	Tidak	Tidak	Tidak	Tidak	Ya	Tidak	Tidak
2	2023/06/05 12:20:04 PM GMT+7	NINGRUM	Perempuan	21	FASILKOM ILMU KOMPUTER	Tidak sama sekali	Ya	Ya	Ya	Ya	...	Ya	Ya	Ya	Ya	Ya	Ya	Ya
3	2023/06/05 12:26:41 PM GMT+7	Zephi Vera Try Khoirunnisa	Perempuan	20	Arsitektur dan Desain	Cukup baik	Ya	Tidak	Tidak	Ya	...	Tidak	Tidak	Ya	Tidak	Ya	Tidak	Tidak
4	2023/06/05 1:31:04 PM GMT+7	Nisa aisyiyah	Perempuan	22	Fk	Baik	Ya	Tidak	Tidak	Ya	...	Tidak	Tidak	Tidak	Ya	Ya	Ya	Tidak
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Gambar 3. Dataset Unlabelling

Tahapan selanjutnya melakukan proses EDA untuk visualisasi data. Hal ini bertujuan untuk mendapatkan *insight* data. Proses EDA yang dilakukan untuk eksplorasi data pada dataset menggunakan heatmap dan bar plot yang ditunjukkan dalam Gambar 4 dan 5. *Heatmap* menjelaskan korelasi antaratribut dataset. Berdasarkan Gambar 4, semakin cerah warna suatu kolom antar 2 atribut, maka semakin kuat hubungan antaratributnya. Namun, jika semakin gelap warna dari suatu kolom antar 2 atribut tersebut, maka semakin lemah pula hubungan antar atributnya.Dari *heatmap* dataset *mental disorder* tersebut, dapat diketahui bahwa salah satu pasangan atribut yang memiliki korelasi kuat adalah atribut *anger* dan *having.nightmares* dengan nilai korelasinya sebesar 0.99. Sedangkan salah satu pasangan atribut yang memiliki korelasi lemah adalah atribut *introvert* dan *close.friend* dengan nilai korelasinya sebesar -0.65.

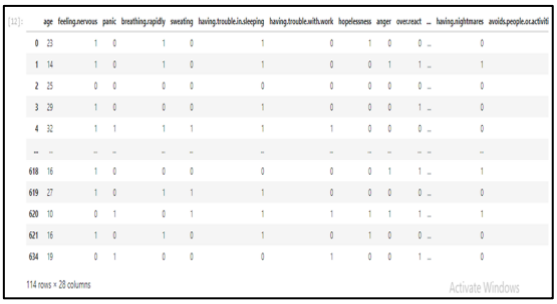


Gambar 4. Visualisasi Heatmap



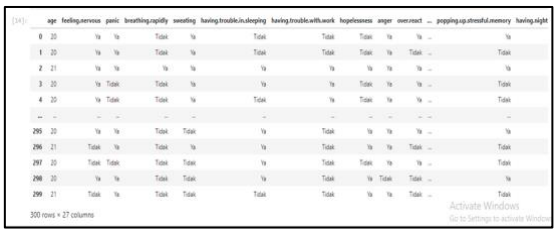
Gambar 5. Bar Plot Disorder

Berdasarkan Gambar 5 dapat diketahui jumlah penderita gangguan kesehatan mental pada tiap jenisnya. Pada Gambar 5 menunjukkan bahwa gangguan kesehatan mental yang paling banyak dialami adalah gangguan kesehatan mental jenis ASD. Sedangkan gangguan kesehatan mental yang jarang dialami adalah gangguan kesehatan mental jenis *psychotic depression*.



Gambar 6. Hasil Pre-processing Dataset Labeling

Berdasarkan beberapa tahapan *pre-processing* yang dilakukan diperoleh hasil yang ditunjukkan pada Gambar 6. Pada Gambar 6 merupakan hasil *pre-processing* data *labeling*. Pada tahap ini, dilakukan pengecekan nama kolom. Oleh karena terdapat kolom yang tidak sesuai penamaannya, maka penamaan kolom tersebut diubah atau disempurnakan. Hal ini bertujuan untuk memudahkan proses analisis. Kemudian, pada data tersebut dilakukan proses deteksi dan *handling missing value*. Berdasarkan hasil pengecekan *missing value*, tidak ditemukan adanya *record* data yang kosong pada setiap kolom. Setelah itu, dilakukan pula deteksi data duplikat. Hal ini bertujuan untuk menghindari adanya redundansi data yang nantinya mempengaruhi proses analisis. Pada pengecekan data duplikat ini, ditemukan banyak sekali data yang redundan. Hingga akhirnya, *record* data *labeling* tersisa 114 *record* data seperti yang ditunjukkan pada Gambar 7.



Gambar 7. Hasil Pre-processing Dataset Unlabeling

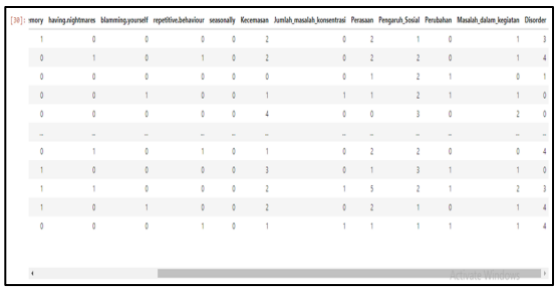
Pada Gambar 7 merupakan hasil *pre-processing* data *unlabeling*. Pada tahap ini dilakukan penghapusan beberapa kolom yang tidak diperlukan dalam proses analisis. Setelah itu dilakukan pula pengubahan nama kolom mengikuti dataset *labelling*. Tahapan selanjutnya adalah *Feature Engineering*. Pada tahap ini, fitur-fitur yang ada pada setiap kolom dianalisis, dan ditemukan bahwa terdapat beberapa fitur yang mempunyai kemiripan atau kesamaan. Oleh karena itu, untuk mengatasinya fitur-fitur ini dijadikan menjadi satu kesatuan. Pada data *labeling* tidak hanya dilakukan penggabungan fitur-fitur saja. Akan tetapi, juga dilakukan pengubahan nama label menjadi lebih pendek lagi. Hal ini bertujuan untuk memudahkan proses analisis. Di samping itu, pada kolom d*isorder* baik pada data labeling maupun unlabeling dilakukan penggabungan untuk gangguan yang hampir sama, yaitu

- MDD dan psychotic deprission: Gangguan Mood,
- ASD dan PDD : Gangguan Perkembangan,
- Loneliness: Gangguan Kepribadian,
- anexiety, PTSD, OCD, bipolar, dan ADHD: Gangguan Kecemasan,
- eating disorder: Gangguan Makan,
- sleeping disorder: Gangguan Tidur.

Berdasarkan penggabungan *disorder* tersebut dari 11 menjadi 6 *disorder*, sehingga perlu dilakukan perubahan juga pada disorder menjadi angka 0-5, yaitu:

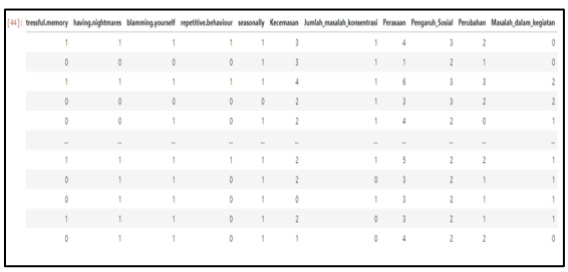
- Gangguan Mood dilambangkan angka 0,
- Gangguan Perkembangan dilambangkan angka 1,
- Gangguan Kepribadian dilambangkan angka 2,
- Gangguan Kecemasan dilambangkan angka 3,
- Gangguan Makan dilambangkan angka 4,
- Gangguan Tidur dilambangkan angka 5.

Berikut adalah hasil tahapan *feature engineering* pada dataset *labeling* seperti pada Gambar 8.



Gambar 8. Hasil Pre-processing Dataset Unlabeling

Sedangkan pada dataset *unlabeling*, juga dilakukan proses penggabungan beberapa fitur dan perpendekan nama kolom. Hal ini juga bertujuan untuk memudahkan proses analisis. Di samping itu, pada dataset *unlabeling* juga dilakukan proses perubahan nilai data. Nilai data yang awalnya "Ya" dan "Tidak", diubah menjadi "1" dan "0". Hasil proses *feature engineering* pada dataset *unlabeling* ditunjukkan pada Gambar 9.



Gambar 9. Hasil Feature Engineering Dataset Unlabeling

Setelah dataset di proses maka menghasilkan data yang bersih, rapi, dan siap digunakan. Langkah selanjutnya adalah tahap pemodelan data. Di tahap ini, algoritma yang digunakan untuk perbandingan pemodelan adalah Regresi Logistik. Dengan menggabungkan data *labeling* yang merupakan *Supervised Learning* dan data *unlabeling* yang merupakan *Unsupervised Learning* sehingga menjadi *Semi-Supervised Learning*. Kemudian, dari penggabungan tersebut data *Semi-Supervised Learning* dibagi menjadi 3 proposisi yaitu

- 1:1 dimana diambil 114 data label dan 100 data *random unlabeling*
- 1:2 diambil 114 data label dan 200 data *random unlabeling*
- 1:3 dimana diambil 114 data label dan 300 data keseluruhan *unlabeling*.

Splitting data tersebut menggunakan perbandingan persentase split data (data training : data testing), yaitu 70:30 dan 75: 25.

Setelah tahap pemodelan data, melakukan evaluasi model dengan menampilkan *matrix coefficient*, akurasi hasil data kombinasi, dan waktu prosesnya. Evaluasi model pertama dilakukan pada split data 70:30 dengan 3 proporsi data yaitu

a. proporsi 1:1
Pada proporsi data 1:1 jumlah kombinasi dataset sejumlah 214 sehingga data training sejumlah 149 dan data testing sejumlah 65. Hasil classification report untuk proporsi 1:1 ditunjukkan pada Gambar 10.

```
Accuracy pada Data Kombinasi: 0.9230769230769231
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        19
           1       1.00      0.75      0.86         4
           2       1.00      0.86      0.92         7
           3       0.81      1.00      0.90        13
           4       0.95      0.86      0.90        22
           5       0.00      0.00      0.00         0

    accuracy                           0.92        65
   macro avg       0.79      0.75      0.76        65
weighted avg       0.95      0.92      0.93        65

Processing Time: 0.038895368576049805 seconds
```

Gambar 10. Hasil classification report untuk proporsi 1:1 (split data 70:30)

b. proporsi 1:2
Pada proporsi data 1:2 jumlah kombinasi dataset sejumlah 314 sehingga data training sejumlah 219 dan data testing sejumlah 95. Hasil classification report untuk proporsi 1:2 ditunjukkan pada Gambar 11.

```
Accuracy pada Data Kombinasi: 0.9473684210526315
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.97      0.99        36
           1       1.00      1.00      1.00         3
           2       1.00      1.00      1.00         5
           3       0.77      0.91      0.83        11
           4       0.94      0.92      0.93        36
           5       1.00      1.00      1.00         4

    accuracy                           0.95        95
   macro avg       0.95      0.97      0.96        95
weighted avg       0.95      0.95      0.95        95

Processing Time: 0.03889966011047363 seconds
```

Gambar 11. Hasil classification report untuk proporsi 1:2 (split data 70:30)

c. proporsi 1:3
Pada proporsi data 1:3 jumlah kombinasi dataset sejumlah 414 sehingga data training sejumlah 289 dan data testing sejumlah 125. Hasil classification report untuk proporsi 1:3 ditunjukkan pada Gambar 12.

```
Accuracy pada Data Kombinasi: 0.968
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99        36
           1       1.00      0.80      0.89         5
           2       1.00      0.85      0.92        13
           3       0.96      1.00      0.98        26
           4       0.95      0.98      0.97        43
           5       1.00      1.00      1.00         2

    accuracy                           0.97       125
   macro avg       0.98      0.94      0.96       125
weighted avg       0.97      0.97      0.97       125

Processing Time: 0.04887080192565918 seconds
```

Gambar 12. Hasil classification report untuk proporsi 1:3 (split data 70:30)

Evaluasi model kedua dilakukan pada split data 75:25 dengan 3 proporsi data yaitu

a. proporsi 1:1
Pada proporsi data 1:1 jumlah kombinasi dataset sejumlah 214 sehingga data training sejumlah 160 dan data testing sejumlah 54. Hasil classification report untuk proporsi 1:1 ditunjukkan pada Gambar 13.

```
Accuracy pada Data Kombinasi: 0.9259259259259259
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        16
           1       1.00      0.75      0.86         4
           2       1.00      1.00      1.00         4
           3       0.77      1.00      0.87        10
           4       1.00      0.85      0.92        20
           5       0.00      0.00      0.00         0

    accuracy                           0.93        54
   macro avg       0.79      0.77      0.77        54
weighted avg       0.96      0.93      0.94        54

Processing Time: 0.07494163513183594 seconds
```

Gambar 13. Hasil classification report untuk proporsi 1:1 (split data 75:25)

b. proporsi 1:2
Pada proporsi data 1:2 jumlah kombinasi dataset sejumlah 314 sehingga data training sejumlah 235 dan data testing sejumlah 79. Hasil classification

report untuk proporsi 1:2 ditunjukkan pada Gambar 14.

```
Accuracy pada Data Kombinasi: 0.9873417721518988
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.97      0.98        32
           1       1.00      1.00      1.00         2
           2       1.00      1.00      1.00         4
           3       1.00      1.00      1.00        10
           4       0.96      1.00      0.98        27
           5       1.00      1.00      1.00         4

    accuracy                           0.99        79
   macro avg       0.99      0.99      0.99        79
weighted avg       0.99      0.99      0.99        79

Processing Time: 0.10668110847473145 seconds
```

Gambar 14. Hasil *Classification Report* untuk Proporsi 1:2 (split data 75:25)

c. proporsi 1:3

Pada proporsi data 1:3 jumlah kombinasi dataset sejumlah 414 sehingga data training sejumlah 10 dan data testing sejumlah 104. Hasil classification report untuk proporsi 1:3 ditunjukkan pada Gambar 15.

```
Accuracy pada Data Kombinasi: 0.9423076923076923
Classification Report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        30
           1       1.00      0.60      0.75         5
           2       1.00      0.82      0.90        11
           3       0.95      0.95      0.95        22
           4       0.92      0.97      0.94        34
           5       1.00      1.00      1.00         2

    accuracy                           0.94       104
   macro avg       0.97      0.89      0.92       104
weighted avg       0.95      0.94      0.94       104

Processing Time: 0.14760446548461914 seconds
```

Gambar 14. Hasil *Classification Report* untuk Proporsi 1:3 (split data 75:25)

Berdasarkan hasil evaluasi model kombinasi data dan split data diperoleh nilai akurasi dan lama proses yang ditunjukkan dalam Tabel 1.

Tabel 1 Hasil Evaluasi dengan Kombinasi Data dan Rasio Split

Algoritma	**Kombinasi Data**	**Split Rasio**	**Akurasi (%)**	**Lama Proses (detik)**
Regresi Logistik	1:1	70:30	92.3	0.04
	1:2		94.74	0.04
	1:3		96.8	0.05
	1:1	75:25	92.59	0.75
	1:2		**98.73**	0.11
	1:3		94.23	0.15

Berdasarkan Tabel 1 didapatkan hasil akurasi terbaik pada algoritma Regresi Logistik dengan kombinasi data *labelling* dan *unlabelling* sebesar 1:2 pada split data 75:25 sebesar 98.73% tetapi lama waktu pemrosesan bukan yang terbaik yaitu selama 0.11 detik. Hal ini dikarenakan waktu yang dibutuhkan lebih lama untuk pemroses data train.

KESIMPULAN

Pada penelitian ini digunakan dua jenis data, yakni data *labelling* dan data *unlabelling* dimana kedua jenis data tersebut digabungkan sehingga menjadi *Semi-Supervised Learning*. Pada data *labelling* digunakan klasifikasi dengan algoritma Regresi Logistik sedangkan pada data *unlabelling* digunakan metode *Pseudo Labeling*. Dan dihasilkan kedua metode tersebut berhasil diterapkan pada penelitian ini. Dari hasil proses analisis dan pemodelan dataset yang telah dilakukan terlihat bahwa penggunaan perbandingan antara data labeling dan unlabeling dapat mempengaruhi hasil akurasi serta lama waktu pemrosesan. Selain itu, penggunaan metode *Pseudo Labeling* dengan algoritma Regresi Logistik sangat cocok untuk studi kasus *mental health* yang kami gunakan karena menghasilkan akurasi mencapai 98.73% dari proporsi data sebesar 114:200 (1:2) dan split data 75:25.

DISEMINASI

Artikel ini telah diseminasikan pada Seminar Nasional Teknologi Informasi dan Komunikasi (SEMNASTIK) APTIKOM Tahun 2024 yang diselenggarakan oleh Universitas Methodist Indonesia pada tanggal 24-26 Oktober 2024.

DAFTAR PUSTAKA

Aloysius, S., & Salvia, N. (2021). Analisis Kesehatan Mental Mahasiswa Perguruan Tinggi Pada Awal Terjangkitnya Covid-19 di Indonesia. *Jurnal Citizenship Virtues*, 83-97.

Alwi, W., Ermawati, & Husain, A. (2018). Analisis Regresi Logistik Biner Untuk Memprediksi Kepuasan Pengunjung Pada Rumah Sakit Umum Daerah Majene. *Jurnal MSA*, 20-26.

Dzil Kamalah, A., & Nafiah, H. (2023). Gejala Mental Emosional dan Upaya dalam Meningkatkan Kesehatan Jiwa Remaja. *Jurnal Keperawatan Berbudaya Sehat*, 2986–8548.

Endriyani, S., Lestari, R. D., Lestari, E., & Napitu, I. C. (2022). Gangguan Mental Emosional Dan Depresi Pada Remaja. *Health Care Nursing Journal*, 429–434.

Florensa, Hidayah, N., Sari, L., Yousrihatin, F., & Litaqia, W. (2023). Gambaran Kesehatan Mental Emosional Remaja (Overview). *Jurnal Kesehatan*, 2721–8007.

Juwairi, K. P. (2019). *Pemanfaatan Teknik Semi-Supervised Learning Untuk Klasifikasi Dokumen Medis.* Yogyakarta: Universitas Islam Indonesia Press.

Kesehatan, K. (2018). *Hasil Utama Riskesdas 2018.* Jakarta: Kementrian Kesehatan Indonesia.

Novitasari, D. A., & dkk. (2019). Analisis Regresi Logistik Ordinal pada Kepuasan Pelanggan Mebel Lamongan. *Jurnal Penelitian Ilmu Manajemen*, 841-848.

Nurhaeni, A., Marisa, D. E., & Oktiany, T. (2022). Peningkatan Pengetahuan Tentang Gangguan Kesehatan Mental Pada Remaja. *JAPRI Jurnal Pengabdian Masyarakat Kesehatan*, 29–34.

Pratama, A., Nurcahyo, A. C., & Firgia, L. (2023). Penerapan Machine Learning dengan Algoritma Logistik Regresi untuk Memprediksi Diabetes. *Jurnal Nasional Corisindo*, 116-121.

Pratiwi, K., & Rusinani, D. (2022). Literatur review : Gangguan mental depresi pada wanita. *Jurnal Ilmu Kebidanan*, 103–110.

Putri, A. W., Wibhawa, B., & Gutama, A. S. (2020). Kesehatan Mental Masyarakat Indonesia (Pengetahuan, Dan Keterbukaan Masyarakat Terhadap Gangguan Kesehatan Mental). *Prosiding KS: Riset & PKM* (pp. 147-300). Sumedang: Departemen Kesejahteraan Sosial Universitas Padjajaran.

Radiani, W. A. (2019). Kesehatan Mental Masa Kini dan Penanganan Gangguannya Secara Islami. *Journal of Islamic and Law Studies*, 87–113.

Rahmawaty, F., Silalahiv, R. P., & Mansyah, B. (2022). Faktor-Faktor Yang Mempengaruhi Kesehatan Mental Pada Remaja. *Jurnal Surya Medika*, 277–281.

Rajoub, B. (2020). Characterization of biomedical signals: Feature engineering and extraction. *in Biomedical Signal Processing and Artificial Intelligence in Healthcare, Elsevier*, 29–50.

Ridlo, I. A. (2020). Pandemi COVID-19 dan Tantangan Kebijakan Kesehatan Mental di Indonesia. *Jurnal Psikologi dan Kesehatan Mental*, 155-164.

Sari, A. P., Prasetya, D. A., Yasuno, T., Sihananto, A. N., Al Haromainy, M. M., & Saputra, W. S. (2022). Forecasting Model of Wind Speed and Direction by Convolutional Neural Network-Deep Convolutional Long Short Term Memory. *2022 IEEE 8th Information Technology of International Seminar (ITIS)* (hal. 200-205). Surabaya: IEEE Xplore.

Sari, A. P., Suzuki, H., Kitajima, T., Yasuno, T., Prasetya, D. A., & Arifuddin, R. (2022). Short-Term Wind Speed and Direction Forecasting by 3DCNN and Deep. *IEEJ Transactions on Electrical and Electronic Engineering*, 1620–1628.