IJAI (Indonesian Journal of Applied Informatics)
Vol. 8 No. 2 Tahun 2024 pISSN: 2548-3846, eISSN: 2598-5981

Deteksi Dini Gejala Stres pada Mahasiswa Berdasarkan Faktor-Faktor Penyebabnya Menggunakan Metode *Logistic Regression*

Auliya Afifah Adnan Hakim*, Rizal Adi Saputra, Statiswaty
Program Studi S1 Teknik Informatika, Universitas Halu Oleo
*Email: auliyaafifah24@gmail.com

Info Artikel

Kata Kunci :
logistic regression, kesehatan mental*, machine learning,* stres

Keywords :
logistic regression, mental health, machine learning, stress

Tanggal Artikel :
Dikirim : 16 Januari 2024
Direvisi : 12 April 2024
Diterima : 30 Mei 2024

Abstrak

Di era saat ini, kesehatan mental seringkali diabaikan dibandingkan dengan kesehatan fisik. Hal ini sangat berlaku bagi mahasiswa yang berjuang dengan tekanan yang meningkat dalam kehidupan kuliah. Oleh karena itu, ada kebutuhan mendesak untuk solusi yang dapat mendukung kesehatan mental, salah satunya adalah metode *Logistic Regression*. Penelitian ini menegaskan efektivitas metode *Logistic Regression* dalam memprediksi dan mendeteksi dini gejala stres berdasarkan faktor penyebabnya. Eksplorasi fitur dan parameter mengungkap variasi tingkat akurasi, mencapai puncak 95%, diikuti oleh 88% dan 61%. Menganalisis hasil menggunakan *Counseling_Service_Use* sebagai output menunjukkan keahlian model dalam memprediksi hasil positif, meskipun dengan kecenderungan untuk memprediksi data negatif sebagai positif, dan sebaliknya. Sementara itu, model yang menggunakan *Chronic_Illness* dan *Stress_Level* sebagai output menunjukkan kinerja luar biasa dalam memprediksi semua kelas. Secara keseluruhan, penelitian ini memberikan dukungan kuat untuk efektivitas *Logistic Regression* dalam memprediksi gejala stres, memperkaya pemahaman tentang penerapannya dalam konteks kesehatan mental. Ini menunjukkan bahwa metode ini dapat digunakan sebagai alat yang efektif untuk deteksi dini gejala stres.

Abstract

In the current era, mental health often takes a backseat compared to physical health. This is particularly true for students who are grappling with the increasing pressures of college life. Therefore, there is an urgent need for solutions that can support mental health, one of which is the Logistic Regression method. This study affirms the effectiveness of the Logistic Regression method in predicting and early detecting symptoms of stress based on their causative factors. Exploration of features and parameters revealed variations in accuracy rates, reaching a peak of 95%, followed by 88% and 61%. Analyzing results using Counseling_Service_Use as the output demonstrated the model's proficiency in predicting positive outcomes, albeit with a tendency to predict negative data as positive, and vice versa. Meanwhile, models employing Chronic_Illness and Stress_Level as outputs exhibited outstanding performance in predicting all classes. Overall, this research provides robust support for the effectiveness of Logistic Regression in predicting stress symptoms, enriching the understanding of its application in the context of mental health. This shows that this method can be used as an effective tool for early detection of stress symptoms.

1. PENDAHULUAN

Saat ini, masyarakat semakin peduli dengan kesehatan fisik mereka, namun kesehatan mental tidak mendapat perhatian yang sama. Sekalipun mereka sadar bahwa mereka menderita penyakit mental kronis, banyak orang memilih untuk tidak mencari pengobatan karena takut dengan apa yang dipikirkan orang lain, keyakinan bahwa mereka sudah gila, ketidaksukaan terhadap dokter, atau ketiganya[1]. Begitupun dengan mahasiswa dihadapkan pada tekanan yang semakin meningkat dalam menjalani kehidupan perkuliahan. Meskipun kesadaran akan pentingnya kesehatan fisik telah menjadi fokus utama masyarakat, kesehatan mental mahasiswa tampaknya sering diabaikan. Pada penelitian sebelumnya kesehatan mental siswa tingkat ketiga menjadi perhatian masyarakat karena adanya kesenjangan antara permintaan akan layanan dan dukungan yang ditawarkan pada tingkat krisis[2]. Oleh karena itu, solusi yang segera diperlukan untuk mengatasi stigma, memberikan dukungan emosional, dan menciptakan lingkungan yang mendukung kesehatan mental. Salah satu pendekatan yang dapat diambil adalah penerapan metode peramalan kesehatan mental menggunakan pendekatan pembelajaran mendalam dan algoritma *machine learning*, seperti metode *Logistic Regression*.

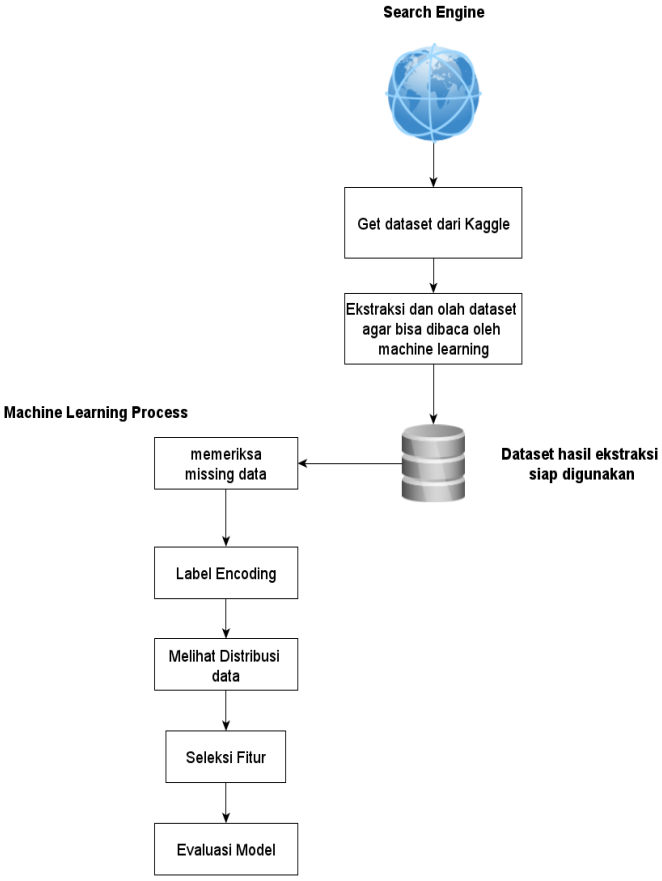
Untuk *case* kesehatan mental, *stress*, *depression*, *anxiety* dan sejenisnya banyak yang menerapkan algoritma *machine learning* untuk mendapatkan hipotesis prediksi yang sesuai dan mengetahui bagaimana cara kerja metode ini serta mengetahui besar akurasinya. Sebagai contoh pada penelitian sebelumnya berjudul *"Prediction of depression among senior citizens using machine learning classifiers"*, pada penelitian ini mereka menerapkan *machine learning* untuk memprediksi *anxiety* dan tingkat depresi pada lansia dengan 10 jenis *classifier* ke dalam dataset berjumlah 510 baris data lansia yang didapatkan dari kar Medical College and Hospital, Kolkata. Dengan menerapkan *ten-fold cross-validation* didapatkan hasil akurasi terbaik model machine learning menggunakan RF sebesar 89%.i [3]. Faktor pendukung lainnya seperti riwayat penyakit kronis, umur, lingkungan dan hubungan dengan orang lain juga mempengaruhi gejala stres pada setiap orang [4].

Penelitian ini menerapkan *machine learning* dengan metode *Logistic Regression* untuk mendeteksi dini gejala stres pada mahasiswa dan menganalisis fitur sehingga dapat diketahui faktor apa saja yang paling mempengaruhi *stress* pada mahasiswa. Pentingnya menangani masalah kesehatan mental di kalangan mahasiswa tidak dapat diabaikan. Oleh karena itu, solusi yang segera diperlukan untuk mengatasi stigma, memberikan dukungan emosional, dan menciptakan lingkungan yang mendukung kesehatan mental. Salah satu pendekatan yang dapat diambil adalah penerapan metode peramalan kesehatan mental menggunakan pendekatan pembelajaran mendalam dan algoritma pembelajaran mesin, seperti *Logistic Regression*. *Logistic Regression* terbukti memberikan akurasi yang lebih tinggi dibandingkan dengan algoritma pembelajaran mesin lainnya, sehingga dapat menjadi solusi efektif dalam deteksi dan mengatasi masalah kesehatan mental di kalangan mahasiswa.

Penelitian ini bertujuan untuk mengidentifikasi dini gejala stres pada mahasiswa dengan mengembangkan metode menggunakan *Logistic Regression*. Metode ini diharapkan dapat mendeteksi dini gejala stres pada mahasiswa, sehingga dapat membantu dalam penanganan dan pencegahan stres lebih lanjut. Penelitian ini juga bertujuan untuk menganalisis faktor-faktor yang mempengaruhi stres mahasiswa. Penelitian ini menyelidiki faktor-faktor seperti riwayat penyakit, umur, lingkungan, dan hubungan interpersonal yang ada dalam *dataset*. Faktor-faktor ini diharapkan dapat dijadikan indikator penting dalam memahami tingkat stres mahasiswa. Dengan demikian, hasil penelitian ini diharapkan dapat memberikan kontribusi penting dalam bidang kesehatan mental, khususnya dalam konteks kesehatan mental mahasiswa.

2. METODE PENELITIAN

Pada tahap ini metode yang digunakan serta alur penelitian tergambar pada Gambar 1. Mulai dari mencari dataset, ekstraksi, memproses di *machine learning* dan terakhir implementasi menggunakan algoritma *Logistic Regression* untuk mendapatkan hasil deteksi dan akurasi yang optimal.



Gambar 1. Alur Penelitian

Pada Gambar 1. penelitian dimulai dari akses data melalui *search engine* seperti *Mozilla Firefox, Google Chrome, Microsoft Edge* dan sejenisnya. Selanjutnya setelah dataset didapatkan, dataset diolah kembali menggunakan *tools Google Colab* karena ada beberapa data yang masih berupa kata yang tidak bisa dibaca oleh program sehingga harus dilakukan *label encoding.* Setelah *dataset* siap digunakan tahap selanjutnya adalah memeriksa data *missing* dan menampilkan beberapa plot grafik untuk melihat distribusi data. Selanjutnya adalah memilih fitur yang berpotensi bisa mendeteksi *stress* dini pada dataset dan terakhir yakni evaluasi model.

2.1 Dataset

Dataset dapat didefinisikan sebagai kumpulan data yang disusun sedemikian rupa untuk mendukung analisis atau percobaan tertentu. *Dataset* biasanya terdiri dari entitas data, variabel, dan catatan yang terkait satu sama lain [5]. *Dataset* yang digunakan pada penelitian ini dikumpulkan melalui survei yang dilakukan oleh Google Forms dari mahasiswa Universitas untuk memeriksa situasi akademik dan kesehatan mental mereka saat ini. Meskipun data ini bukan berasal dari penelitian langsung oleh penulis. *Dataset* ini memberikan para ilmuwan data, analis, dan para profesional terkait kesempatan untuk berpartisipasi dalam tantangan dan proyek berbasis data untuk memecahkan masalah dunia nyata. *Dataset* ini juga mendukung kolaborasi dan pembelajaran di komunitas ilmu data. *Dataset* bernama *student_mental_health_survey* dengan total jumlah data sebanyak 7023 baris dan 20 kolom yang terdiri dari beberapa fitur diantaranya, *Age, Course, Gender, CGPA, Stress_Level, Depression_Score, Anxiety_Score, Sleep_Quality, Physical_Activity, Diet_Quality, Social_Support, Relationship_Status, Substance_Use, Counseling_Service_Use, Family_History, Chronic_Illness, Financial_Stress, Extracurricular_Involvement, Semester_Credit_Load, Residence_Type.* Baru-baru ini istilah *big data* menjadi sangat populer di seluruh dunia. Selama beberapa tahun terakhir, *big data* telah mulai memasuki sistem pelayanan kesehatan[6]. Data mentah yang didapat dari *search engine* ini akan diolah kembali dan diekstraksi agar bisa di proses di *machine learning model.* Variabel data mentah yang digunakan antara lain pada Tabel 1 berikut.

Tabel 1. Variabel Penelitian

No	*Variabel*	*Keterangan*
1	*Age*	*Umur*
2	*Course*	*Prodi yang diambil individu*
3	*Gender*	*Jenis kelamin*
4	*CGPA*	*Indeks Prestasi Kumulatif individu*
5	*Stress_Level*	*Tingkat Stres yang dialami individu*
6	*Depression_Score*	*Skor yang menggambarkan tingkat depresi yang dialami individu*
7	*Anxiety_Score*	*Skor yang menggambarkan tingkat kecemasan yang dialami individu*
8	*Sleep_Quality*	*Kualitas tidur yang dialami individu*
9	*Physical_Activity*	*Tingkat aktivitas fisik*
10	*Diet_Quality*	*Kualitas pola makan individu*
11	*Social_Support*	*Tingkat dukungan sosial yang diterima individu*
12	*Relationsip_Status*	*Status hubungan individu*
13	*Substance_Use*	*Frekuensi penggunaan zat, seperti alkohol atau rokok*
14	*Counseling_Service_Use*	*Penggunaan layanan konseling*
15	*Family_History*	*Riwayat kesehatan mental dalam keluarga individu*
16	*Chronic_Illness*	*Penyakit Kronis yang dimiliki individu*
17	*Financial_Stress*	*Tingkat stres keuangan yang dialami individu*
18	*Extracurricular_Involvement*	*Keterlibatan individu dalam kegiatan ekstrakurikule*
19	*Semester_Credit_Load*	*Jumlah SKS yang diambil individu dalam semester tersebut*
20	*Residence_Type*	*Tipe tempat tinggal individu*

2.2 Implementasi pada Google Colab

Ada beberapa proses yang dilakukan pada saat akan mengimplementasikan dataset dan menguji metode *Logistic Regression* menggunakan platform Google Colab. Google Colab (Colaboratory) adalah platform daring yang disediakan oleh Google untuk melakukan pengolahan dan analisis data menggunakan Python. Colab memberikan lingkungan pengembangan berbasis cloud yang memungkinkan pengguna untuk membuat, menjalankan, dan membagikan *notebook* Jupyter tanpa memerlukan konfigurasi atau instalasi perangkat lunak di mesin lokal. Colab memanfaatkan kekuatan komputasi awan Google, termasuk penggunaan unit pemrosesan grafis (GPU) secara gratis, yang dapat mempercepat proses pelatihan model *machine learning* dan komputasi berat lainnya [7] . Berikut adalah metode yang digunakan pada saat implementasi.

2.2.1 Memeriksa *Missing Data*

Missing data atau data yang hilang merupakan area yang menonjol ketika berhadapan dengan analisis data. Nilai-nilai yang hilang ini biasanya disebabkan oleh kesalahan manusia saat memproses data, kesalahan mesin karena tidak berfungsinya peralatan, penolakan responden untuk menjawab pertanyaan tertentu, putus sekolah dan menggabungkan data yang tidak terkait [8]. Strategi pengelolaan *missing data* lainnya melibatkan mencari tetangga terdekat di antara vektor fitur yang lengkap dan menggantikan nilai fitur yang hilang dengan nilai dari tetangga terdekat ini. Pendekatan ini dapat menjadi kurang efektif ketika terdapat sejumlah besar contoh dengan peningkatan jumlah fitur yang hilang [9].

2.2.2 *Label Encoding*

Label Encoding adalah teknik pemberian representasi numerik atau angka pada nilai-nilai kategori dalam suatu fitur. *Label encoding* memungkinkan penggunaan pengklasifikasi *biner* dalam jumlah yang dapat disesuaikan tergantung pada kuantisasi, pengkodean, dan fungsi *decoding*. Hal ini membuka kemungkinan untuk meningkatkan akurasi masalah regresi dengan ruang desain besar yang mencakup fungsi kuantisasi, pengkodean, *decoding*, dan kesalahan[10].

2.2.3 Melihat Distribusi Data

Pada tahap ini dilakukan pemrosesan pada *machine learning* yang merujuk pada pola sebaran atau penyebaran nilai-nilai dalam dataset. Ini mencakup bagaimana data terdistribusi pada berbagai kelas atau kelompok, serta frekuensi kemunculan nilai-nilai tertentu dalam atribut-atribut dataset. Distribusi dataset yang baik akan memberikan gambaran yang seimbang dan representatif dari berbagai kondisi atau kategori yang ada dalam dataset [11]. Parameter pengukur distribusi data pada dataset ini adalah melihat dari sebaran umur dan keterkaitan dengan faktor yang mempengaruhi stres.

2.2.4 Seleksi Fitur

Dalam *machine learning*, seleksi fitur adalah proses pemilihan subset fitur dari kumpulan data yang tersedia untuk digunakan dalam pengembangan model. Motivasi utama dari pemilihan fitur melibatkan pencarian fitur yang memberikan kontribusi maksimal terhadap kinerja model, menghasilkan wawasan yang berharga terkait data, dan potensial penghematan dalam pengumpulan atau pemrosesan data. Pemilihan fitur mendapat perhatian signifikan dalam penelitian analisis data karena dampak positifnya terhadap kualitas model[12].

2.2.5 Evaluasi Model

Logistic regression adalah suatu metode regresi yang digunakan untuk menganalisis hubungan antara hasil biner atau kategorik dengan beberapa faktor yang mempengaruhi. Metode ini mencakup variasi seperti regresi logistik berganda, regresi logistik bersyarat, regresi logistik politom, regresi logistik ordinal, dan regresi logistik kategorik yang berdekatan [13]. Variabel yang ingin kita prediksi dikenal sebagai variabel dependen tunggal dan beberapa variabel independent [14]. Beberapa matrik evaluasi umum termasuk akurasi (*accuracy)*, presisi (*precision*), sensitivitas (*recall*), *F1-score*, dan area di bawah kurva ROC (AUC-ROC) [13]. Pada tahap ini dataset diimplementasikan dengan beberapa model kemudian dibandingkan untuk mencari performa model terbaik untuk mendeteksi atau memprediksi gejala stres pada mahasiswa. *Logistic regression* adalah algoritma klasifikasi yang digunakan untuk memprediksi probabilitas kejadian suatu peristiwa. Algoritma ini sering digunakan untuk tugas klasifikasi *biner*, di mana output yang diprediksi adalah salah satu dari dua kategori[15].

Proses ini dimulai dengan inisialisasi bobot dan bias. Setiap fitur dalam dataset memiliki bobot yang terkait dengannya, dan model juga memiliki bias yang memungkinkan penyesuaian terhadap pergeseran. Selanjutnya, dilakukan komputasi linear, yaitu menghitung nilai linear dari kombinasi bobot dan nilai fitur. Setelah itu, dilakukan transformasi logistik atau *sigmoid*. Fungsi logistik digunakan untuk mengkonversi nilai linear menjadi probabilitas. Kemudian, digunakan fungsi *loss* seperti *Cross-Entropy* untuk mengukur perbedaan antara nilai prediksi dan nilai aktual.

Selanjutnya, dilakukan optimasi parameter. Proses ini melibatkan proses optimisasi, seperti gradien turun, untuk meminimalkan fungsi loss. Bobot dan bias diperbarui untuk meningkatkan akurasi model. Akhirnya, setelah model telah dilatih, model tersebut dapat digunakan untuk membuat prediksi baru pada data yang belum terlihat. Dengan demikian, model *logistic regression* ini dapat digunakan untuk mendeteksi dini gejala stres pada mahasiswa.

Model regresi logistik adalah kasus khusus dari model linier umum. Ini terdiri dari prediktor linier pada persamaan (1) berikut :

$$\eta = \beta 0 + \beta 1 x 1 + \ldots + \beta m x m \quad (1)$$

Keterangan:
η = Prediksi
$\beta 0$ = Konstanta regresi
$\beta 1, \beta 2, \ldots, \beta m$ = Koefisien regresi
$x1, x2, \ldots, xm$ = Variabel bebas

Fungsi tautan yang menghubungkan probabilitas bersyarat ke prediktor linier pada persamaan (2) berikut:

$$P(y|x) = \mu(\eta) / f(\eta) \quad (2)$$

Keterangan:
$P(y|x)$ = Probabilitas bersyarat dari peristiwa y dengan syarat bahwa peristiwa x telah terjadi
$\mu(\eta)$ = Fungsi kepadatan probabilitas dari peristiwa y di titik η
$f(\eta)$ = Fungsi kepadatan probabilitas dari peristiwa x di titik η

Fungsi tautan adalah logaritma peluang, disebut *logit* diberikan pada persamaan (3) berikut :

$$\text{logit} = \log \mu/1 - \mu \,.\, \text{MLE}\, \hat{\beta} \quad (3)$$

Keterangan:
μ = Probabilitas atau proporsi yang ditransformasikan oleh fungsi *logit*
$\text{MLE}\, \hat{\beta}$ = Perkiraan parameter beta yang diperoleh dengan Estimasi Maximum Likelihood

yang diperoleh dengan meminimalkan logaritma negatif dari fungsi kemungkinan [16] dalam persamaan (4) berikut :

$$\min_{\beta} L(\beta) = - \sum_{i=1}^{n} \{y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)\} \quad (4)$$

3. HASIL DAN PEMBAHASAN

3.1 *Missing Data Check*

Pada tahap ini dilakukan pengecekan data *null* atau missing menggunakan *library is.null()* untuk tiap kolom dataset. Berikut adalah hasilnya.

Tabel 2. *Missing Data*

Missing Data Check		
Kolom	Total	Percent
Substance_Use	15	0.002136
CGPA	12	0.001709
Relationship_Status	0	0.000000
Age	0	0.000000
Financial_Stress	0	0.000000
Family_History	0	0.000000
Social_Support	0	0.000000
Diet_Quality	0	0.000000
Stress_Level	0	0.000000
Stress_Level	0	0.000000
Physical_Activity	0	0.000000
Depression_Score	0	0.000000
Chronic_Illness	0	0.000000

Berdasarkan hasil pada Tabel 2. didapatkan bahwa hanya ada 2 kolom yang memiliki *missing data* yakni *Substance_Use* dan *CGPA* dengan total masing-masing 12-15 dengan persentase 0.002136 dan 0.001709. Untuk menghilangkan *missing data* tersebut kita perlu melakukan *drop data* yang *missing* agar tidak terjadi kesalahan nanti pada saat implementasi model.

3.2 *Label Encoding*

Pada tahap ini menggunakan *library from sklearn.preprocessing import LabelEncoder* untuk mengubah kolom dengan nilai yang masih berbentuk kata atau huruf ke dalam bentuk nilai angka sesuai dengan beberapa jenis data pada masing-masing kolom. Hasil dari proses *label encoding* tertera pada hasil berikut.

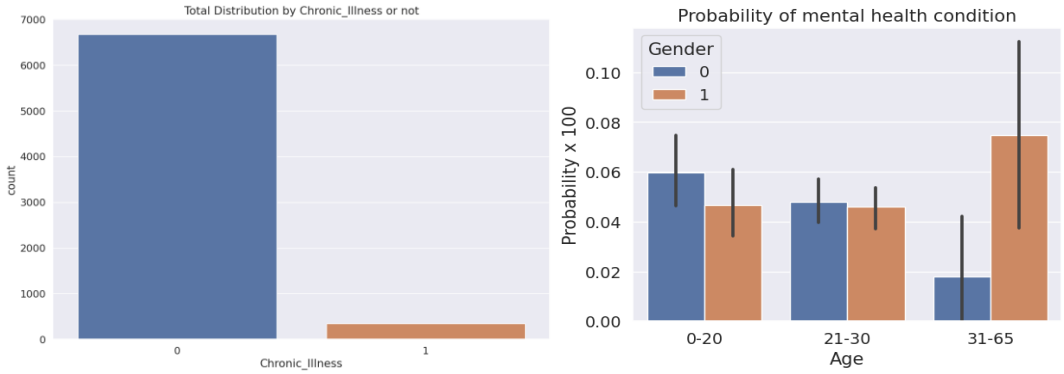
label_Age [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35]
label_Course [0, 1, 2, 3, 4, 5]
label_Gender [0, 1]
label_Stress_Level [0, 1, 2, 3, 4, 5]
label_Depression_Score [0, 1, 2, 3, 4, 5]
label_Anxiety_Score [0, 1, 2, 3, 4, 5]
label_Sleep_Quality [0, 1, 2]
label_Physical_Activity [0, 1, 2]
label_Diet_Quality [0, 1, 2]
label_Social_Support [0, 1, 2]
label_Relationship_Status [0, 1, 2]
label_Counseling_Service_Use [0, 1, 2]
label_Family_History [0, 1]
label_Chronic_Illness [0, 1]
label_Financial_Stress [0, 1, 2, 3, 4, 5]
label_Extracurricular_Involvement [0, 1, 2]
label_Semester_Credit_Load [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]
label_Residence_Type [0, 1, 2]
label_age_range ['0-20', '21-30', '31-65']

Pada hasil terlihat bahwa tiap kolom sudah berganti menjadi angka pada nilainya, seperti contoh *Gender* yang terdiri dari *Male* dan *Female* berubah label menjadi angka 0 dan 1. Begitupun seterusnya nilai index label dimulai dari 0 - banyaknya jenis label yang ada pada kolom *dataset*.

3.3 Distribusi Data

Selanjutnya untuk mengetahui penyebaran *dataset* dan agar mengetahui fitur atau kolom faktor mana yang paling mempengaruhi langkah selanjutnya adalah dengan melakukan analisis distribusi data. Ada beberapa parameter yang digunakan pada tahap ini yakni, *Age, Chronic Illness,* dan melihat *Probabilitas mental_health* berdasarkan *Gende*r. Berikut adalah hasilnya.

Gambar 2. Distribusi Data Berdasarkan kolom *stress_level* dan *Age*

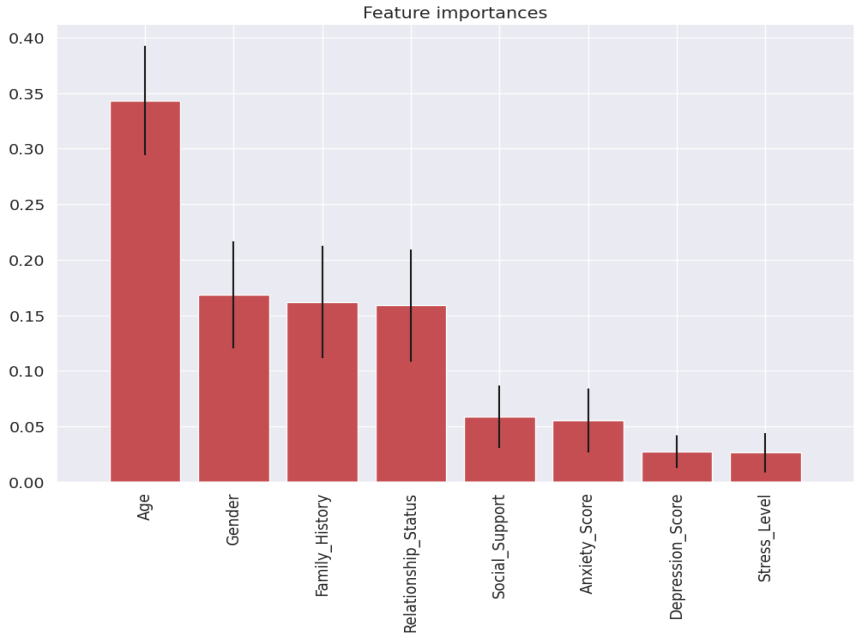


Gambar 3. Distribusi Data Berdasarkan *Chronic_Illness* dan *Gender*

Pada Gambar 2. terlihat bahwa rentang umur mahasiswa yang mengalami gejala *stress* berdasarkan *stress_level* pada kolom yang terdiri dari 4 *level stress*. Karena data *age* atau umur sudah dilakukan *label encoding* tadi sehingga rentang data umur hanya 0-20 saja. Pada Gambar 3. terjadi kenaikan di rentang umur 0-10 yang berarti umur aktualnya adalah 18 - 28 tahun. Selanjutnya pada Gambar 4. digambarkan bahwa distribusi data berdasarkan yang memiliki penyakit kronis adalah >6000 baris data adalah tidak memiliki penyakit kronis dan >1000 memiliki penyakit kronis. Lalu pada probabilitas *mental health condition* berdasarkan *gender* adalah yang paling banyak memiliki gejala *stress* wanita ditandai dengan warna *orange* pada rentang usia 31-65 tahun. Untuk laki-laki yang paling banyak memiliki gejala *stress* pada rentang usia 0-20 tahun.

3.4 Seleksi Fitur

Tahap selanjutnya adalah seleksi fitur, pada tahap ini diharapkan dapat ditemukan fitur mana saja yang berpotensi untuk menjadi parameter pada saat menjalankan model machine learning dengan metode *logistic regression*. Setelah dilakukan seleksi fitur dengan menggunakan *library forest.feature_importances_* didapatkan hasil pada Gambar 4. Pada hasil analisis terlihat pada gambar diagram batang dibawah ini bahwa age atau umur merupakan salah satu fitur yang paling penting untuk menjadi parameter pengukuran model untuk deteksi dini gejala *stress* pada mahasiswa. Fitur selanjutnya adalah *Gender, Family_history,* dan *relationship status* yang memiliki nilai yang sama. Selanjutnya disusul dengan *social_support, anxiety_score, depression_score* dan *stress level* fitur yang bisa digunakan untuk mengukur akurasi deteksi dini pada penerapannya menggunakan metode *Logistic Regression*.



Gambar 4. Hasil Seleksi Fitur

3.5 Evaluasi Model

Tahap terakhir pada penelitian ini adalah melakukan evaluasi terhadap model atau metode algoritma yang digunakan. Berikut adalah hasil evaluasi model menggunakan algoritma atau metode *Logistic Regression* dengan beberapa kombinasi fitur.

1. Hasil evaluasi model dengan kombinasi fitur

Tabel 3. Evaluasi Model

feature_cols	*X*	*y*	*Akurasi Model*
['Age', 'Gender', 'Family_History', 'Relationship_Status', 'Social_Support', 'Anxiety_Score', 'Depression_Score', 'Stress_Level']	*df[feature_cols]*	*df.Counseling_Service_ Use*	61%
	df[feature_cols]	*df.Chronic_Ilness*	95%
	df[feature_cols]	*df.Stress_level*	88%

2. *Confusion Matrix Stress_Level*

Tabel 4. *Confusion Matrix Stress_Level*

Actual	*Predicted*
0	329
1	353
2	407
3	416
4	305
5	297

Berdasarkan nilai-nilai tersebut, dapat disimpulkan bahwa:

1. Model memprediksi dengan benar 329 data yang sebenarnya bernilai 1.
2. Model memprediksi dengan benar 353 data yang sebenarnya bernilai 2.
3. Model memprediksi dengan benar 407 data yang sebenarnya bernilai 3.

Untuk nilai 1,2,3 merupakan label dari Stres, Non Stres dan NaN. Dengan demikian, model memiliki akurasi sebesar: (329 + 353 + 407) / (329 + 0 + 0 + 353 + 0 + 0 + 407 + 0) = 0.88. Output dari model ini dapat dilihat pada tabel 5.

Tabel 5. *Output Predict Model*

	Index	*Stress or not*	*Label*
0	1815	5	*Stress*
1	2955	5	*Stress*
2	1615	2	*nan*
3	3318	3	*Stress*

4	2570	1	*Non stress*
…	…	…	…
2102	6795	5	*Stress*
2103	789	3	*Stress*
2104	561	5	*Stress*
2105	298	5	*Stress*
2106	3873	4	*Stress*

4. KESIMPULAN

Berdasarkan penelitian ini, dapat disimpulkan bahwa metode *Logistic Regression* terbukti efektif untuk prediksi dan deteksi dini gejala stres berdasarkan faktor penyebabnya. Eksplorasi fitur dan parameter menunjukkan variasi dalam tingkat akurasi, dengan nilai tertinggi mencapai 95%, diikuti oleh 88% dan 61%. Hasil analisis menggunakan *Counseling_Service_Use* sebagai nilai *output* menunjukkan bahwa model memiliki kinerja yang baik dalam memprediksi data positif, namun masih terdapat kecenderungan untuk memprediksi data negatif sebagai positif, dan sebaliknya. Sementara itu, model yang menggunakan *Chronic_Illness* dan *Stress_Level* sebagai output menunjukkan kinerja yang sangat baik dalam memprediksi semua kelas. Keseluruhan, penelitian ini memberikan dukungan kuat untuk keefektifan *Logistic Regression* dalam konteks prediksi gejala stres.

DAFTAR PUSTAKA

[1] Jage, S., Chaudhari, S., Jatte, M., Mhatre, A., & Mane, V. (2023). "Predicting Mental Health Illness using Machine Learning." In 2023 3rd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-5).

[2] M. Hill, N. Farrelly, C. F. Clarke, dan M. Cannon, "Student mental health and well-being: Overview and Future Directions," Irish Journal of Psychological Medicine, vol. 1-8, 2020.

[3] Usman, M., Haris, S. H., & Fong, A. C. (2020). "Prediction of Depression using Machine Learning Techniques: A Review of Existing Literature." In 2020 IEEE 2nd International Workshop on System Biology and Biomedical Systems (SBBS) (pp. 1-3).

[4] C. S. Wahyuningsih, A. A. Subijanto, dan B. Murti, "Logistic Regression on Factors Affecting Depression among the Elderly," Journal of Epidemiology and Public Health, 2019.

[5] Smith, J., & Johnson, A. (2020). "CIFAR-10: Canadian Institute For Advanced Research Dataset." [Image Dataset]. CIFAR. https://www.cs.toronto.edu/~kriz/cifar.html

[6] A. N. Haq, A. Khattak, N. Jamil, M. A. Naeem, dan F. Mirza, "Data Analytics in Mental Healthcare," Sci. Program., vol. 2020, pp. 2024160:1-2024160:9, 2020.

[7] Google Colab. "Google Colab: An Easy Way to Learn and Use Python." [Online]. Available: https://colab.research.google.com/ , 29 December 2023.

[8] T. Emmanuel et al., "A survey on missing data in machine learning," Journal of Big Data, vol. 8, 2021.

[9] D. P. Mesquita, J. P. Gomes, and A. H. Souza, "A Minimal Learning Machine for Datasets with Missing Values," in International Conference on Neural Information Processing, 2015.

[10] D. Shah, Z. Xue, dan T. M. Aamodt, "Label Encoding for Regression Networks," ArXiv, vol. abs/2212.01927, 2022.

[11] Johnson, R., & Smith, T. (2020). "Understanding Dataset Distribution in Machine Learning." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2001-2015.

[12] P. Cunningham, B. Kathirgamanathan, dan S. Delany, "Feature Selection Tutorial with Python Examples," ArXiv, vol. abs/2106.06437, 2021.

[13] Q. Wang, S. Yu, X. Qi, Y. Hu, W. J. Zheng, J. Shi, dan H. Yao, "Overview of logistic regression model analysis and application," Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine], vol. 53, no. 9, pp. 955-960, 2019.

[14] W. D. Suryono, W. L. Pratitis, Y. Asmara, and A. Wahyudi, "Multi Variable Regresi Sebagai Prediksi Area Terdampak Kebakaran Hutan," IJAI (Indonesian Journal of Applied Informatics), vol. 6, no. 2, pp. 121-126, 2022.

[15] A. Deb, B. Samadder, S. Chowdhury, S. Das, dan S. Banarjee, "Measuring Mental Health Condition using Logistic Regression," International Journal of Engineering Technology and Management Sciences. [Dalam Penerbitan], 2023.

[16] S. Kost, O. Rheinbach, dan H. Schaeben, "Logistic regression for potential modeling," PAMM, vol. 19, 2019.