Klasifikasi Depresi Menggunakan Regresi Logistik Dan K-Nearest Neighbor Berdasarkan Faktor Demografis Dan Kesehatan

Hana Ramadila[1], Elisa Desi Syafitri[2], Susandri Susandri[3], Ahmad Zamsuri[4]
[1,2,3,4] Magister Ilmu Komputer Sekolah Pascasarjana Universitas Lancang Kuning
[1,2,3,4] Jl. Yos Sudarso KM. 8 Rumbai, Pekanbaru, Riau, telp. 0811 753 2015
e-mail: [1]hanaramadila22@gmail.com , [2]elisadesisyafitri@gmail.com

Abstrak

Depresi merupakan gangguan kesehatan mental yang berdampak besar terhadap kualitas hidup dan masih memiliki prevalensi tinggi di Indonesia, sementara upaya deteksi dini sering terkendala oleh keterbatasan alat skrining yang praktis dan akurat. Penelitian ini bertujuan untuk mengembangkan model klasifikasi depresi berbasis variabel demografis dan kesehatan dengan menggunakan algoritma Regresi Logistik dan K-Nearest Neighbor (KNN), serta membandingkan kinerjanya dalam memprediksi gejala depresi. Penelitian dilakukan secara eksperimental menggunakan dataset publik berjumlah 413.768 entri dari Kaggle yang telah melalui tahap preprocessing, analisis matriks korelasi, pelatihan model, dan evaluasi menggunakan metrik akurasi dan confusion matrix. Hasil menunjukkan bahwa regresi logistik memperoleh akurasi sebesar 73,16% pada data pelatihan dan 72,72% pada data pengujian, sedangkan KNN hanya mencapai 66,98%. Analisis korelasi juga mengidentifikasi hubungan bermakna antara status pekerjaan dengan pendapatan serta antara usia dengan tingkat aktivitas fisik sebagai prediktor penting depresi. Dapat disimpulkan bahwa regresi logistik lebih unggul dibandingkan KNN untuk klasifikasi depresi pada dataset ini, sehingga berpotensi diimplementasikan dalam sistem skrining dini depresi. Penelitian lanjutan disarankan menambahkan variabel psikososial dan algoritma lain guna meningkatkan performa model.
***Kata Kunci**: Depresi, Regresi Logistik, K-Nearest Neighbor, Klasifikasi, Kesehatan Mental*

Abstract

Depression is a mental health disorder that significantly affects quality of life and remains highly prevalent in Indonesia, while early detection efforts are often hindered by the lack of practical and accurate screening tools. This study aims to develop a depression classification model based on demographic and health variables using the Logistic Regression and K-Nearest Neighbor (KNN) algorithms, as well as to compare their performance in predicting depressive symptoms. The study employed an experimental approach using a public dataset consisting of 413,768 entries obtained from Kaggle, which underwent preprocessing, correlation matrix analysis, model training, and evaluation using accuracy and confusion matrix metrics. The results show that logistic regression achieved an accuracy of 73.16% on the training data and 72.72% on the testing data, while KNN only reached 66.98%. Correlation analysis also identified meaningful relationships between employment status and income, and between age and physical activity level, as important predictors of depression. It can be concluded that logistic regression outperforms KNN for depression classification on this dataset, making it potentially applicable for early depression screening systems. Future studies are suggested to incorporate additional psychosocial variables and other machine learning algorithms to further improve model performance.
***Keywords**: Depression, Logistic Regression, K-Nearest Neighbor, Classification, Mental Health.*

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi (TIK) yang pesat telah mengubah paradigma pendidikan secara global. Penggunaan teknologi digital kini menjadi kebutuhan mendasar untuk menciptakan pembelajaran yang interaktif, fleksibel, dan efektif [1]. Di Indonesia, implementasi Kurikulum Merdeka menuntut sekolah untuk memfasilitasi peserta didik dengan akses terhadap sumber belajar digital yang luas [2]. Namun, kesenjangan infrastruktur seringkali menjadi penghambat utama dalam pemerataan kualitas pendidikan, terutama di daerah berkembang.

Depresi adalah gangguan kesehatan mental yang signifikan secara global dan lokal yang berdampak pada kesejahteraan fisik dan psikologis individu. Menurut laporan WHO, depresi menyebabkan gangguan fungsional yang tinggi dan merupakan salah satu penyebab utama disabilitas dunia [1]. Di Indonesia, meskipun berbagai upaya telah dilakukan dalam deteksi dan pengobatan kesehatan mental, prevalensi gejala depresi masih cukup tinggi terutama pada kelompok rentan seperti remaja, orang tua, dan tenaga kesehatan; sayangnya, banyak kasus yang tidak terdeteksi karena stigma, kurangnya akses ke layanan, serta keterbatasan alat skrining yang praktis dan efisien.

Salah satu masalah utama dalam pendeteksian depresi adalah bahwa banyak penelitian menggunakan pendekatan tradisional atau satu algoritma saja terutama regresi logistik tanpa mempertimbangkan metode pembelajaran mesin lainnya yang dapat menangkap pola nonlinier dan interaksi antar variabel yang mungkin lebih kompleks. Di sisi lain, penelitian lokal sering hanya menggunakan variabel demografis saja atau variabel psikososial tertentu, dan minim penggunaan variabel kesehatan fisik (seperti penyakit kronis, kualitas tidur, gangguan aktivitas fisik, status kesehatan umum) secara simultan. Hal ini menjadi masalah karena faktor-faktor kesehatan tersebut dalam banyak studi internasional terbukti signifikan dalam mempengaruhi risiko depresi.

Selain itu, sebagian besar penelitian lokal belum membandingkan secara sistematis antara regresi logistik dan K-Nearest Neighbor (KNN) dalam konteks variabel demografis dan kesehatan. Perbandingan tersebut penting karena regresi logistik memberikan keunggulan dalam hal interpretabilitas dan kemudahan implementasi, sementara KNN sebagai metode nonparametrik dapat lebih fleksibel dalam menangani kompleksitas data apabila variabel prediktor memiliki hubungan nonlinier atau distribusi tidak normal. Evaluasi performa kedua metode tersebut dalam metrik seperti akurasi, sensitivitas, spesifisitas, area under ROC curve (AUC) juga belum banyak dilakukan di populasi Indonesia secara lokal.

Penelitian terdahulu yang relevan antara lain “Factors associated with depressive symptoms among adolescents in Indonesia” yang menggunakan data survei nasional (Indonesian Family Life Survey) dan melakukan analisis regresi logistik terhadap variabel demografis dan kesehatan seperti jenis kelamin, pendidikan, status ekonomi, penyakit kronis, kualitas tidur, kebiasaan merokok, dan tipe kepribadian; penelitian tersebut menemukan bahwa sekitar 29.1% remaja melaporkan gejala depresi, dengan variabel seperti jenis kelamin wanita, kualitas tidur rendah, dan penyakit kronis secara signifikan berkaitan dengan risiko depresi [2]. Penelitian “Determinants of Depression in Indonesian Youth (RISKESDAS 2018)” juga menggunakan regresi logistik multivariat untuk mengidentifikasi faktor risiko di kalangan remaja dan dewasa muda (15-24 tahun), menunjukkan bahwa jenis kelamin wanita, merokok, konsumsi alkohol, memiliki penyakit kronis, serta depresi orang tua adalah faktor risiko yang signifikan [3]. Studi “Depression among Older Adults in Indonesia: Prevalence, Role of Health-Related Factors” memanfaatkan data IFLS-5 dan analisis regresi logistik untuk orang tua (≥ 60 tahun), menemukan prevalensi depresi sebesar 16.3% dan faktor yang signifikan meliputi status sosial ekonomi subjektif, kesehatan diri yang dianggap buruk, ketergantungan aktivitas sehari-hari, insomnia, dan jatuh sebagai indikator penting [4]. Ada juga penelitian nasional “Determinant factors related to stress, resilience, and depression among health workers during the COVID-19 pandemic in Indonesia” yang menerapkan regresi logistik dan menemukan sejumlah faktor demografis dan pekerjaan (jumlah anak, status pekerjaan, tanggungan keluarga, usia) yang berkaitan dengan gejala depresi pada tenaga kesehatan di fasilitas rujukan COVID-19 [5].

Meskipun demikian, dari kajian pustaka di atas, ditemukan bahwa belum banyak penelitian di Indonesia yang menggunakan KNN dalam model klasifikasi depresi, apalagi membandingkannya secara langsung dengan regresi logistik dalam satu dataset yang mencakup variabel demografis dan kesehatan secara menyeluruh. Sebagai contoh,

penelitian "Komparasi Kinerja Algoritma Random Forest, Decision Tree, Naïve Bayes, dan KNN dalam Prediksi Tingkat Depresi Mahasiswa menggunakan Student Depression Dataset" menunjukkan bahwa KNN adalah salah satu dari beberapa algoritma yang diuji, namun tidak secara khusus diperbandingkan dengan regresi logistik dalam konteks variabel kesehatan fisik dan determinan lokal lainnya [5]. Penelitian "Deteksi Dini Gangguan Kesehatan Mental berdasarkan Sentimen dan Data Terstruktur" di Jurnal Sistemasi juga membandingkan beberapa metode termasuk regresi logistik, namun fokusnya lebih pada data terstruktur dan aspek demografis/sentimen daripada variabel kesehatan fisik / medis secara mendalam [6].

Berdasarkan identifikasi masalah tersebut, penelitian ini bertujuan untuk mengisi kekosongan tersebut dengan fokus pada dua tujuan utama: pertama, mengembangkan model klasifikasi depresi berdasarkan variabel demografis dan kesehatan menggunakan regresi logistik dan KNN dalam dataset lokal; dan kedua, mengevaluasi serta membandingkan kinerja kedua model tersebut dalam hal akurasi, sensitivitas, spesifisitas, dan AUC, agar bisa diketahui metode mana yang paling efektif untuk prediksi dini depresi dalam konteks Indonesia.

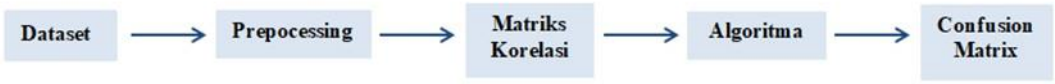
Secara teori, penelitian ini didasari pada pendekatan biopsikososial, yang menyatakan bahwa depresi bukan hanya hasil dari faktor psikologis saja, tetapi juga faktor biologis (kesehatan fisik, penyakit kronis), sosial (status ekonomi, dukungan sosial), dan demografis (usia, jenis kelamin, pendidikan, pekerjaan). Dalam kerangka ini, regresi logistik digunakan sebagai metode parametik untuk memodelkan logit probabilitas depresi berdasarkan kombinasi linier dari variabel prediktor, memungkinkan interpretasi langsung dari koefisien dan odds ratio; sedangkan KNN dipilih karena kemampuannya sebagai metode non-parametrik yang dapat menangkap pola nonlinieritas antar fitur dan interaksi kualitas tidur, komorbiditas, jenis kelamin, umur, dan faktor kesehatan lainnya tanpa membuat asumsi distribusi variabel.

Kebaruan dari penelitian ini terletak pada beberapa aspek: pertama, penggunaan dataset lokal Indonesia yang mencakup variabel demografis dan kesehatan fisik secara komprehensif (termasuk penyakit kronis, kualitas tidur, status kesehatan diri, aktivitas fisik) dalam model klasifikasi depresi; kedua, perbandingan langsung antara regresi logistik dan KNN dalam satu studi dengan evaluasi menyeluruh terhadap performa model (akurasi, sensitivitas, spesifisitas, AUC); ketiga, hasil penelitian diharapkan memberi interpretabilitas yang kuat koefisien regresi logistik akan menunjukkan pengaruh variabel, sedangkan KNN akan diuji terhadap pengaruh parameter k dan scaling fitur, sehingga bisa memberikan rekomendasi praktis bagi kebijakan kesehatan mental lokal; dan keempat, penelitian ini diharapkan menjadi dasar bagi pengembangan sistem skrining dini depresi yang efisien dan dapat dioperasikan di lapangan, terutama di fasilitas kesehatan primer dan institusi pendidikan.

2. METODE PENELITIAN

2.1. Tahapan Penelitian

Penelitian yang dilakukan merupakan hasil dari eksperimen yang bertujuan membandingkan hasil dari algoritma pada data klasifikasi depresi. Tahapan penelitian yang digunakan dapat dilihat sebagai berikut pada Gambar 1.



Gambar 1. Tahapan Penelitian

Dataset adalah data mentah dalam bentuk tabel yang akan diolah dengan algoritma. Penelitian ini menggunakan data publik yang dapat diakses untuk semua

orang. Data didapat dari website www.kaggle.com. Dataset yang digunakan yaitu "depression_data" berjumlah 413.768 data. Hasil dataset dapat ditemukan pada Gambar 2, yang mencakup beragam atribut data. Tabel tersebut memberikan gambaran awal tentang kompleksitas data yang dikumpulkan, yang kemudian menjadi subjek utama dalam proses analisis dan evaluasi model klasifikasi.

```
[15]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 413768 entries, 0 to 413767
Data columns (total 16 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Name                        413768 non-null  object
 1   Age                         413768 non-null  int64
 2   Marital Status              413768 non-null  object
 3   Education Level             413768 non-null  object
 4   Number of Children          413768 non-null  int64
 5   Smoking Status              413768 non-null  object
 6   Physical Activity Level     413768 non-null  object
 7   Employment Status           413768 non-null  object
 8   Income                      413768 non-null  float64
 9   Alcohol Consumption         413768 non-null  object
 10  Dietary Habits              413768 non-null  object
 11  Sleep Patterns              413768 non-null  object
 12  History of Mental Illness   413768 non-null  object
 13  History of Substance Abuse  413768 non-null  object
 14  Family History of Depression 413768 non-null  object
 15  Chronic Medical Conditions  413768 non-null  object
dtypes: float64(1), int64(2), object(13)
memory usage: 50.5+ MB
```

Gambar 2. Dataset "depression_data"

2.2 Preprocessing Data

Preprocessing adalah proses pengolah data atau gambar asli sebelum data atau gambar tersebut diproses oleh algoritma yang diusulkan pada metode penelitian. Preprocessing diperlukan dalam proses penemuan pengetahuan karena data yang berkualitas diperlukan untuk menghilangkan data yang tidak ada atau NULL [7]. Adapun preprocessing yang telah dilakukan pada Gambar 3.

```
[6]: df = pd.read_csv('depression_data.csv')
     df
```

[6]:

	Name	Age	Marital Status	Education Level	Number of Children	Smoking Status	Physical Activity Level	Employment Status	Income
0	Christine Barker	31	Married	Bachelor's Degree	2	Non-smoker	Active	Unemployed	26265.67
1	Jacqueline Lewis	55	Married	High School	1	Non-smoker	Sedentary	Employed	42710.36
2	Shannon Church	78	Widowed	Master's Degree	1	Non-smoker	Sedentary	Employed	125332.79
3	Charles Jordan	58	Divorced	Master's Degree	3	Non-smoker	Moderate	Unemployed	9992.78
4	Michael Rich	18	Single	High School	0	Non-smoker	Sedentary	Unemployed	8595.08
...	...	...	...	...	...	...	...	...	...
413763	Sean Miller	68	Married	Master's Degree	0	Former	Moderate	Employed	109233.43
413764	Christina Brown	26	Single	Bachelor's Degree	0	Current	Active	Employed	96760.97
413765	Matthew Jenkins	57	Married	Bachelor's Degree	0	Non-smoker	Sedentary	Employed	77353.26
413766	Gary Faulkner	71	Married	Associate Degree	2	Non-smoker	Sedentary	Unemployed	24557.08
413767	Joseph Johnson	62	Widowed	Master's Degree	0	Former	Moderate	Employed	107125.74

413768 rows × 16 columns

Gambar 3. Preprocessing Data

2.3 Matriks Korelasi

Matriks korelasi adalah tabel dengan koefisien korelasi untuk berbagai variabel. Matriks menunjukkan bagaimana semua pasangan nilai yang mungkin ada dalam tabel saling terkait satu sama lain. Matriks ini sangat berguna untuk menemukan pola dalam kumpulan data besar dan meringkasnya.

Sering kali ditampilkan dalam bentuk tabel, dengan baris dan kolom untuk setiap variabel dan koefisien korelasi untuk setiap pasangan variabel yang ditulis dalam setiap sel. Koefisien korelasi berkisar antara -1 dan +1. Berikut cara kerja matriks :

a) Nilai 1 menunjukkan hubungan yang kuat dan positif antara dua variabel.
b) Nilai 0 menunjukkan bahwa tidak ada hubungan antara keduanya.
c) Nilai -1 menunjukkan hubungan yang kuat dan negatif atau terbalik.

2.4 Algoritma

Algoritma pemrograman adalah dasar dari semua aktivitas pemrograman dan digunakan untuk menyelesaikan masalah. Adapun algoritma yang digunakan dalam penelitian ini yaitu Regresi Logistik dan K-Nearest Neighbor (KNN).

1. Regresi Logistik

Regresi Logistik adalah komponen dari teknik data mining yang digunakan untuk menganalisis data. Regresi logistik membantu menjelaskan hubungan antara satu variabel respons atau variabel dependen, dengan satu atau lebih variabel prediktor [8]. Dalam hal, klasifikasi depresi, algoritma Regresi Logistik memiliki keunggulan dan kelemahan. Memiliki tujuan untuk meningkatkan pemahan kinerja algoritma, variabel yang mempengaruhi kinerja, serta kondisi ideal untuk penerapannya. Adapun model dari algoritma Regresi Logistik dapat dilihat pada Gambar 4 sebagai berikut :

Regresi Logistik

```
[256]: scaler = MinMaxScaler()
       X = scaler.fit_transform(X_selected)

[259]: X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size=0.1, shuffle = True, random_state=42)

[262]: model = LogisticRegression()

[265]: X_train.shape

[265]: (372391, 7)

[268]: model.fit(X_train, Y_train)
```

Gambar 4. Regresi Logistik

K- Nearest Neighbor (KNN) adalah metode untuk mengklasifikasikan objek dengan data pembelajaran yang memiliki jarak paling dekat atau perbedaan nilai paling kecil dengan objek yang diuji, jumlah tetangga terdekat ditentukan secara manual dan ditunjukkan dengan k . Dalam hal, klasifikasi depresi, algoritma K- Nearest Neighbor (KNN) memiliki keunggulan dan kelemahan. Memiliki tujuan untuk meningkatkan pemahan kinerja algoritma, variabel yang mempengaruhi kinerja, serta kondisi ideal untuk penerapannya. Satu metrik digunakan untuk menghitung seberapa jauh jarah antara data uji dan data latih.

Data yang digunakan menentukan nilai k terbaik. Meskipun nilai k tinggi umumnya akan mengurangi efek noise, batasan antar klasifikasi akan menjadi lebih tidak jelas. Adapun model dari Algoritma K-Nearest Neighbor (KNN) dapat dilihat pada Gambar 5.

Gambar 5. Model K-Nearest Neighbor (KNN)

2. Confusion Matrix

Confusion matrix adalah jumlah data uji yang diklasifikasikan dengan benar dan salah, memudahkan evaluasi akurasi sistem klasifikasi [9]. Confusion matrix digunakan untuk mengevaluasi tingkat akurasi proses klasifikasi yang telah dilakukan. Tingkat akurasi ini menunjukkan proporsi jumlah prediksi yang benar. Empat komponen utama confusion matrix menunjukkan kinerja model dan dapat dilihat pada Tabel 1.

d) True Positive (TP) : ini terjadi ketika model dengan benar memprediksi sampel sebagai positif (kelas yang diinginkan).
e) b) False Positive (FP) : ini terjadi ketika model salah memprediksi sampel sebagai positif, padahal seharusnya negatif (kelas yang tidak diinginkan).
f) c) True Negative (TN) : ini terjadi ketika model dengan benar memprediksi sampel sebagai negatif.
g) d) False Negative (FN) : ini terjadi ketika model salah memprediksi sampel sebagai negatif, padahal seharusnya positif.

Tabel 1. Confusion Matrix

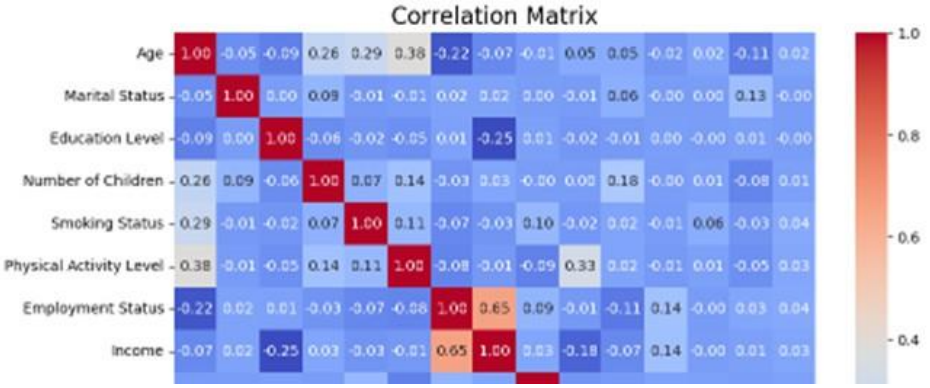
	Nilai Aktual	
	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

3. HASIL DAN PEMBAHASAN

Berikut ini adalah hasil yang dari penelitian yang telah dilakukan, dapat dilihat sebagai berikut :

3.1 Correlation Matrix

Berdasarkan hasil correlation matrix yang didapat dapat dilihat pada Gambar 6 sebagai berikut :



Gambar 6. Correlation Matrix

Analisis korelasi dilakukan untuk mengidentifikasi hubungan antar variabel demografis dan kesehatan yang digunakan sebagai prediktor dalam model klasifikasi.

Hasil visualisasi matriks korelasi menunjukkan bahwa sebagian besar pasangan variabel memiliki nilai korelasi yang relatif rendah, yaitu berada pada rentang antara -0,20 hingga 0,30. Hal ini menandakan bahwa hubungan linear antar variabel prediktor secara umum lemah.

Beberapa pasangan variabel menunjukkan hubungan yang lebih menonjol, antara lain korelasi positif sedang antara employment status dan income sebesar 0,65. Korelasi ini mengindikasikan bahwa individu dengan status pekerjaan yang lebih stabil cenderung memiliki tingkat pendapatan yang lebih tinggi. Selain itu, ditemukan korelasi negatif sedang antara age dan physical activity level sebesar -0,38 yang menunjukkan bahwa semakin bertambah usia, tingkat aktivitas fisik individu cenderung menurun. Korelasi positif lemah juga tampak antara marital status dan number of children sebesar 0,26.

Tidak ditemukan adanya korelasi yang sangat tinggi ($r > 0,80$) antar variabel, sehingga dapat disimpulkan bahwa data tidak mengandung multikolinearitas yang signifikan. Dengan demikian, seluruh variabel layak digunakan dalam tahap pemodelan tanpa perlu dilakukan penghapusan fitur berdasarkan pertimbangan korelasi.

Bahwa "Korelasi yang cukup kuat antara Employment Status dan Income dapat menjadi indikator penting karena faktor ekonomi diketahui memengaruhi risiko depresi." Sejalan dengan penelitian yang telah dilakukan di Indonesia [10], [11]. Dan "Korelasi negatif antara *Age* dan *Physical Activity Level* juga penting, karena tingkat aktivitas fisik yang rendah pada usia lanjut sering dikaitkan dengan peningkatan risiko gangguan kesehatan mental termasuk depresi." Sejalan dengan penelitian yang telah dilakukan [12].

3.2 Algoritma

Pengujian algoritma penting dilakukan untuk memastikan bahwa metode yang digunakan dapat menghasilkan hasil yang akurat, efisien, dan dapat diandalkan. Dalam kontak mengklasifikasikan depresi digunakan algoritma regresi logistik dan K-Nearest Neighbor (KNN), sebagai berikut :

1 . Regresi Logistik

Pada hasil regresi logistik terdapatnya 2 data yaitu data pelatihan dan data pengujian. Data pelatihan juga dikenal sebagai training data yaitu data yang digunakan untuk melatih model memahami hubungan antara fitur (independen) dan target (dependen). Sementara Data pengujian juga dikenal sebagai testing data yaitu data yang digunakan untuk menguji model yang telah dilatih dengan data pelatihan, guna mengetahui apakah model dapat digeneralisasi dengan baik. Adapun hasil dari data pelatihan yang telah dilakukan dapat dilihat pada Gambar 7.

```
[273]: model.score(X_train, Y_train)

[273]: 0.7315590333815801
```

Gambar 7. Pelatihan

Dari gambar diatas dapat dikatakan bahwa model.score(X_train, Y_train) menghasilkan nilai 0,73155. Nilai ini menunjukkan akurasi model pada data pelatihan. Artinya, model mampu memprediksi sekitar 73,16% dari data pelatihan. Sementara itu hasil dari data pengujian yang telah dilakukan dapat dilihat pada Gambar 8.

```
[278]: y_pred = model.predict(X_test)

[282]: print(accuracy_score(y_pred, Y_test))
       0.7271914348551127
```

Gambar 8. Pengujian

Dari gambar diatas bahwa akurasi pada data uji (accuracy_score(y_pred, Y_test) menghasilkan nilai 0,72719. Bahwa model dapat memprediksi sekitar 72,72% dari data pengujian dengan benar.

Hasil akurasi regresi logistik pada data pelatihan sebesar 73,16% dan data pengujian sebesar 72,72%, menunjukkan bahwa algoritma ini cukup mampu memprediksi kategori depresi tergantung pada faktor yang digunakan.

2. K-Nearest Neighbor (KNN)

Pada hasil KNN ini didapatkannya akurasi. Akurasi adalah metrik evaluasi yang digunakan untuk mengukur seberapa baik model memprediksi label kelas pada data uji. Akurasi memberikan proporsi prediksi yang benar terhadap jumlah total prediksi. Adapun hasil akurasi yang telah dilakukan dapat dilihat pada Gambar 9.

```
[293]: y_pred_knn = knn.predict(X_test)

[296]: accuracy = accuracy_score(Y_test, y_pred_knn)
       print(accuracy)
       0.669816564758199
```

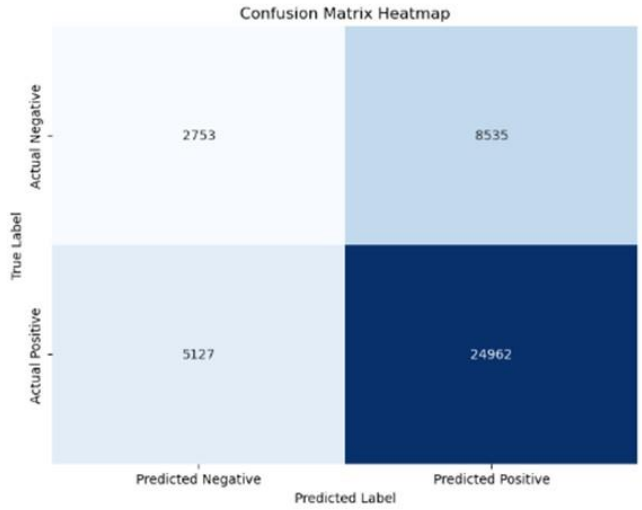
Gambar 9. Akurasi

Fungsi accuracy_score dari scikit-learn digunakan untuk menghitung akurasi model dengan membandingkan label sebenarnya (Y_test) dan hasil prediksi (y_pred_knn). Hasil akurasi yang ditampilkan adalah 0,6698% atau sekitar 66,98%. Nilai akurasi sekitar 66,98% ini dapat memprediksi model sesuai dengan label sebenarnya. Akurasi 66,98% berarti model berhasil memprediksi depresi atau tidak depresi dengan benar pada 66,98% dari sampel data uji. Dalam kasus ini dapat dikatakan bahwa :

a) Y_test adalah label sebenarnya (misalnya, apakah seseorang benar-benar mengalami depresi)
b) y_pred_knn adalah hasil prediksi model, maka akurasi menunjukkan seberapa sering model sesuai dengan kenyataan.

3. Confusion Matrix

Dalam konteks klasifikasi depresi, confusion matrix memberikan pemahaman yang penting tentang bagaimana model bekerja untuk menentukan apakah seseorang mengalami depresi atau tidak depresi, dengan menampilkan dua kelas yaitu kelas positif (individu yang benar-benar mengalami depresi) dan kelas negatif (individu yang tidak mengalami depresi). Namun, untuk memahami lebih lanjut, kita harus mempertimbangkan banyak faktor yang mempengaruhi hasil model. Adapun hasil dari confusion matrix yang telah dilakukan dapat dilihat pada Gambar 10.



Gambar 10. Confusion Matrix

Hasil confusion matrix tersebut didapatkan bahwa nilai TN (True Negative) sebesar 2.753, FP (False Positive) sebesar 8.535, FN (False Negative) sebesar 5.127, dan TP (True Positive) sebesar 24.962.

1. TN (True Negative = 2.753) bahwa individu yang tidak mengalami depresi dan diklasifikasikan dengan benar oleh model. Model mengidentifikasi orang yang sehat dan baik, tetapi jumlah TN yang sedikit menunjukkan bahwa model mungkin lebih cenderung mengidentifikasi orang sebagai depresi (bias terhadap kelas positif).
2. FP (False Positive = 8.535) bahwa individu yang sebenarnya tidak mengalami depresi tetapi model mengidentifikasinya sebagai depresi. Penyebab yaitu overlap gejala dengan penyakit lain seperti kecemasan atau kelelahan yang berlangsung lama. Jawaban yang salah dalam kuesioner atau survei. Dampak dari orang-orang tersebut mungkin diberi diagnosis yang salah atau disarankan untuk menerima perawatan yang tidak diperlukan, yang dapat meningkatkan kecemasan atau menyebabkan stigma
3. FN (False Negative = 5.127) bahwa individu yang mengalami depresi tetapi tidak diidentifikasi oleh model. Penyebab yaitu gejala depresi yang tidak jelas atau anomali (misalnya depresi disertai dengan hiperaktivitas), data yang tidak lengkap atau akurat, serta variasi dalam ekspresi gejala antar individu (misalnya perbedaan gejala laki-laki dan perempuan). Dampak dari Individu ini mungkin tidak mendapatkan perawatan yang mereka butuhkan, sehingga dapat memperburuk kondisi kesehatan mental mereka. Ini merupakan risiko yang signifikan.

TP (True Positive = 24.962) bahwa individu yang mengalami depresi dan diidentifikasi oleh model. Model berfungi dengan baik untuk mengidentifikasi gejela depresi umum seperti suasana hati rendah, gangguan tidur, dan kelelahan. True Positive akan lebih tinggi jika gejala depresi yang dilaporkan cukup jelas dan data yang digunakan berkualitas tinggi, seperti survei kesehatan mental atau diagnosa klinis.

4. KESIMPULAN

Pengembangan model klasifikasi depresi dengan menggunakan algoritma Regresi Logistik dan K-Nearest Neighbor (KNN) mampu memprediksi gejala depresi berdasarkan faktor-faktor demografis dan kesehatan secara cukup baik. Hasil pengujian menunjukkan bahwa model regresi logistik memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan model KNN, yakni sebesar 73,16% pada data pelatihan dan 72,72% pada data pengujian, sedangkan model KNN hanya mencapai akurasi 66,98%. Perbedaan ini menunjukkan bahwa regresi logistik lebih sesuai untuk mengklasifikasikan depresi pada dataset yang digunakan karena kemampuannya dalam menginterpretasikan hubungan linier antar variabel prediktor, sementara KNN tampak kurang optimal pada data berskala besar dan heterogen seperti pada penelitian ini.

Selain itu, analisis matriks korelasi mengungkapkan adanya hubungan yang cukup berarti antara status pekerjaan dengan pendapatan serta hubungan negatif antara usia dan tingkat aktivitas fisik, yang keduanya merupakan indikator penting dalam memprediksi risiko depresi. Temuan ini sejalan dengan teori biopsikososial yang menjadi dasar penelitian, di mana faktor biologis, sosial, dan demografis berperan bersama dalam memengaruhi kesehatan mental seseorang. Ke depan, penelitian ini dapat dikembangkan lebih lanjut dengan memperluas variasi algoritma pembelajaran mesin yang digunakan, menambahkan fitur-fitur psikososial lainnya (misalnya tingkat stres, dukungan sosial, atau kondisi lingkungan), serta melakukan pengujian pada data lapangan di fasilitas kesehatan dan institusi pendidikan agar model yang dihasilkan lebih aplikatif dan dapat diintegrasikan ke dalam sistem skrining depresi berskala nasional. Upaya ini diharapkan mampu mendukung deteksi dini serta penanganan kesehatan mental secara lebih efektif di Indonesia.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Universitas Lancang Kuning, pihak SMP Negeri 8 Kandis, dosen pembimbing, serta semua pihak yang telah membantu dalam pelaksanaan kegiatan penelitian ini.

DAFTAR PUSTAKA

[1] S. Sulidah, T. A. Sugiyatmi, F. Efendi, I. A. Susanti, and A. Bushy, "Determinant factors related to stress, resilience, and depression among health workers during the COVID-19 pandemic in Indonesia," *Electron. J. Gen. Med.*, vol. 21, no. 2, 2024, doi: 10.29333/ejgm/14484.

[2] H. Idris and F. Tuzzahra, "Factors associated with depressive symptoms among adolescents in Indonesia: A cross-sectional study of results from the Indonesia Family Life Survey," *Malaysian Fam. Physician*, vol. 18, pp. 1–9, 2023, doi: 10.51866/oa.265.

[3] I. Y. Suryaputri, R. Mubasyiroh, S. Idaiani, and L. Indrawati, "Determinants of Depression in Indonesian Youth: Findings from a Community-based Survey," *J. Prev. Med. Public Heal.*, vol. 55, no. 1, pp. 88–97, 2022, doi: 10.3961/JPMPH.21.113.

[4] Y. S. Handajani, E. Schröder-Butterfill, E. Hogervorst, Y. Turana, and A. Hengky, "Depression among Older Adults in Indonesia: Prevalence, Role of Chronic Conditions and Other Associated Factors," *Clin. Pract. Epidemiol. Ment. Heal.*, vol. 18, no. 1, pp. 1–10, 2022, doi: 10.2174/17450179-v18-e2207010.

[5] Ismail Setiawan, I. Fatah Yasin, and Y. Tri Desianti, "Komparasi Kinerja Algoritma Random Forest, Decision Tree, Naïve Bayes, dan KNN dalam Prediksi Tingkat Depresi Mahasiswa menggunakan Student Depression Dataset," *J. Ilmu Komput. dan Teknol.*, vol. 6, no. 1, pp. 47–58, 2025, doi: 10.35960/ikomti.v6i1.1756.

[6] P. W. Simarmata and P. T. Prasetyaningrum, "Development of a Student Depression Prediction Model Based on Machine Learning with Algorithm Performance Evaluation," *J. Inf. Syst. Informatics*, vol. 7, no. 2, pp. 1283–1305, 2025, doi: 10.51519/journalisi.v7i2.1087.

[7] A. E. Satriatama *et al.*, "Analisis Klaster Data Pasien Diabetes untuk Identifikasi Pola dan Karakteristik Pasien," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 5, no. 3, pp. 172–182, 2023, doi: 10.47233/jteksis.v5i3.828.

[8] F. R. Suprihati, "Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression," *J. Sist. Cerdas*, vol. 4, no. 3, pp. 155–160, 2021, doi: 10.37396/jsc.v4i3.166.

[9] R. Nurhidayat and K. E. Dewi, "Penerapan Algoritma K-Nearest Neighbor Dan Fitur Ekstraksi N-Gram Dalam Analisis Sentimen Berbasis Aspek," *Komputa J. Ilm. Komput. dan Inform.*, vol. 12, no. 1, pp. 91–100, 2023, doi: 10.34010/komputa.v12i1.9458.

[10] L. Barrass *et al.*, "The association between socioeconomic position and depression or suicidal ideation in low- and middle-income countries in Southeast Asia: a systematic review and meta-analysis," *BMC Public Health*, vol. 24, no. 1, 2024, doi: 10.1186/s12889-024-20986-9.

[11] M. Arizal and I. D. G. K. Wisana, "Mental Health Effects on Job Retention in Indonesia," *J. Dev. Econ.*, vol. 8, no. 1, pp. 1–20, 2023, doi: 10.20473/jde.v8i1.37445.

[12] R. W. Basrowi *et al.*, "Exploring Mental Health Issues and Priorities in Indonesia Through Qualitative Expert Consensus," *Clin. Pract. Epidemiol. Ment. Heal.*, vol. 20, no. 1, pp. 1–9, 2024, doi: 10.2174/0117450179331951241022175443.