

# Implementasi Algoritma Pembelajaran Mesin dalam Diagnosis Depresi pada Individu

**Raden Jiwa Bumi Prajasantana**  
School of Information Science  
and Technology  
Universitas Pelita Harapan  
Tangerang, Indonesia  
01082220020@student.uph.edu

**Nicholas Andrew Tanubrata**  
Faculty of Science and  
Technology  
Universitas Pelita Harapan  
Tangerang, Indonesia  
01112220017@student.uph.edu

**Jonathan Tiong**  
School of Information Science  
and Technology  
Universitas Pelita Harapan  
Tangerang, Indonesia  
01082220017@student.uph.edu

## I. PENDAHULUAN

Menurut *World Health Organization International* (WHO), depresi adalah gangguan mental umum yang mempengaruhi jutaan orang di seluruh dunia, yang digambarkan sebagai suasana hati atau perasaan yang depresif dalam jangka waktu yang panjang [1]. Pada 2019, diperkirakan 280 juta orang di seluruh dunia (3.8% populasi) mengalami depresi, termasuk 5% orang dewasa dan 5.7% orang yang berusia diatas 60 tahun [2]. Depresi berlangsung hampir sepanjang hari dan setiap hari, selama minimal dua minggu. Kondisi ini dapat mempengaruhi semua aspek kehidupan, termasuk masalah atau konflik di sekolah atau tempat kerja, serta hubungan dengan keluarga dan teman. Selain itu, seseorang yang mengalami depresi cenderung mengalami gangguan tidur, perubahan nafsu makan, perasaan rendah diri dan keputusasaan tentang masa depan, serta mengalami kesulitan konsentrasi.

Depresi disebabkan oleh interaksi yang kompleks antara faktor sosial, psikologis, ekonomi, dan biologis. Individu yang mengalami kejadian negatif dalam kehidupannya, seperti pengangguran, krisis keuangan, trauma, kehilangan, dan masalah hubungan interpersonal, memiliki risiko lebih tinggi untuk mengalami depresi [3]. Depresi juga sering berhubungan erat dengan kondisi kesehatan fisik. Faktor-faktor seperti kurangnya aktivitas fisik, gangguan tidur, merokok, dan konsumsi alkohol berlebihan dapat meningkatkan risiko depresi dan memicu masalah kesehatan lainnya [4].

Diagnosis psikiatri merupakan aspek krusial dalam menentukan langkah penanganan yang tepat untuk depresi, serta memastikan bahwa setiap keputusan atau intervensi didasarkan pada bukti ilmiah yang kuat [5]. Dalam mendukung diagnosis yang lebih cepat dan akurat, teknologi seperti *Machine Learning* dapat menjadi alat potensial dalam

membantu profesional kesehatan mental. *Machine Learning*, atau pembelajaran mesin, adalah cabang dari kecerdasan komputasional yang mampu menganalisis pola kompleks dalam data. Melalui pembelajaran mesin, model dapat dilatih untuk mendeteksi tanda-tanda awal depresi dan memprediksi kondisi kesehatan mental serta fisik seseorang berdasarkan data medis sebelumnya [6]. Salah satu studi yang dilakukan oleh Jason Coder Jia menunjukkan penerapan Machine Learning dengan algoritma *Random Forest* dalam memprediksi gangguan mental. Ia telah menggunakan 400,000 *records* dengan 14 variabel berdasarkan demografi, ekonomi, gaya hidup, dan riwayat medis untuk melakukan klasifikasi, dengan model yang mencapai akurasi 70% [7]. Studi lain yang dilakukan oleh Sadik dalam menerapkan Machine Learning, khususnya dengan algoritma *LightGBM* serta algoritma *Boosting* lainnya untuk memprediksi gangguan mental. Penelitian ini menggunakan *dataset* yang sama seperti studi yang dilakukan oleh Jason Coder dan melakukan pemilihan fitur sehingga berhasil mencapai akurasi sekitar 78%. Hal ini menunjukkan bahwa algoritma *Boosting*, seperti *LightGBM* dan *XGBoost* memiliki kemampuan yang efektif dalam menangani deteksi kesehatan mental, termasuk dalam menangani ketidakseimbangan kelas [8].

Hasil ini menggambarkan potensi pembelajaran mesin sebagai solusi dalam mendukung diagnosis yang lebih cepat dan berbasis data dalam konteks kesehatan mental.

Seiring dengan kemajuan teknologi ini, penting untuk memahami lebih dalam mengenai berbagai algoritma pembelajaran mesin yang dapat diterapkan dalam penelitian ini. Metode-metode dalam *supervised machine learning*, algoritma pembelajaran mesin yang menggunakan data-data berlabel untuk memprediksi nilai atau kelas

tertentu, yang relevan untuk mendiagnosis depresi adalah *Support Vector Machine* (SVM), *Random Forest*, dan *Extreme Gradient Boosting* (XGBoost) [9].

Dengan demikian, penelitian ini bertujuan untuk mendiagnosis depresi dengan menerapkan tiga algoritma *supervised machine learning*, serta membandingkan efektivitasnya berdasarkan tingkat akurasi hasil diagnosis. Faktor-faktor yang akan dianalisis meliputi data demografi, ekonomi, gaya hidup, dan riwayat medis. Algoritma *supervised learning* yang terdiri dari *Support Vector Machine* (SVM), *Random Forest*, dan *Extreme Gradient Boosting* (XGBoost) akan dievaluasi dengan mengukur akurasi dan validitas klinis dari hasil diagnosis depresi.

## II. DESKRIPSI DATA

Penelitian ini mengolah *dataset* [10] yang berisi hasil survei terhadap lebih dari 196,851 responden unik dengan berbagai fitur yang berkaitan dengan demografi, status sosial ekonomi, dan gaya hidup, sebagai berikut:

1. **Nama:** Nama lengkap individu, bertipe data objek.
2. **Usia:** Usia individu dalam tahun , bertipe data numerik.
3. **Status Perkawinan:** Status perkawinan individu. Nilai yang mungkin termasuk Lajang, Menikah, Bercerai, dan Janda/Duda, bertipe data objek.
4. **Tingkat Pendidikan:** Tingkat pendidikan tertinggi yang dicapai oleh individu. Nilai yang termasuk adalah: Sekolah Menengah Atas, Gelar Associate, Gelar Sarjana, Gelar Magister, dan Doktor, bertipe data objek.
5. **Jumlah Anak:** Jumlah anak yang dimiliki individu, bertipe data objek.
6. **Status Merokok:** Menunjukkan apakah individu tersebut seorang perokok atau bukan. Nilai yang mungkin adalah Perokok, Mantan Perokok, dan Bukan Perokok, bertipe data objek.
7. **Tingkat Aktivitas Fisik:** Tingkat aktivitas fisik yang dilakukan oleh individu. Nilai yang mungkin termasuk Tidak Banyak Bergerak, Sedang, dan Aktif, bertipe data objek.

8. **Status Pekerjaan:** Status pekerjaan individu. Nilai yang mungkin termasuk Bekerja dan Menganggur, bertipe data objek.
9. **Gajian:** Gajian tahunan individu dalam USD, bertipe data numerik.
10. **Konsumsi Alkohol:** Tingkat konsumsi alkohol. Nilai yang mungkin termasuk Rendah, Sedang, dan Tinggi. Kebiasaan Diet: Kebiasaan diet individu. Nilai yang mungkin termasuk Sehat, Sedang, dan Tidak Sehat, bertipe data objek.
11. **Pola Tidur:** Kualitas tidur. Nilai yang mungkin termasuk Baik, Cukup, dan Buruk, bertipe data objek.
12. **Riwayat Penyakit Mental:** Apakah individu memiliki riwayat penyakit mental. Nilai yang mungkin adalah Ya dan Tidak, bertipe data objek.
13. **Riwayat Penyalahgunaan Zat:** Apakah individu memiliki riwayat penyalahgunaan zat. Nilai yang mungkin adalah Ya dan Tidak, bertipe data objek.
14. **Riwayat Depresi Keluarga:** Menunjukkan apakah ada riwayat depresi dalam keluarga. Nilai yang mungkin adalah Ya dan Tidak, bertipe data objek.
15. **Kondisi Medis Kronis:** Apakah individu memiliki kondisi medis kronis. Nilai yang mungkin adalah Ya dan Tidak, bertipe data objek.

Contoh data yang akan diolah adalah sebagai berikut:

	Name	Age	Marital Status	Education Level	Number of Children	Smoking Status
0	Christine Barker	31	Married	Bachelor's Degree	2	Non-smoker
1	Jacqueline Lewis	55	Married	High School	1	Non-smoker
2	Shannon Church	78	Widowed	Master's Degree	1	Non-smoker
3	Charles Jordan	58	Divorced	Master's Degree	3	Non-smoker
	Physical Activity Level	Employment Status	Income	Alcohol Consumption	Dietary Habits	
	Active	Unemployed	26265.67	Moderate	Moderate	
	Sedentary	Employed	42710.36	High	Unhealthy	
	Sedentary	Employed	125332.79	Low	Unhealthy	
	Moderate	Unemployed	9992.78	Moderate	Moderate	

Sleep Patterns	History of Mental Illness	History of Substance Abuse	Family History of Depression	Chronic Medical Conditions
Fair	Yes	No	Yes	Yes
Fair	Yes	No	No	Yes
Good	No	No	Yes	No
Poor	No	No	No	No

### III. STUDI PUSTAKA

#### A. Support Vector Machine (SVM)

*Support Vector Machine* (SVM) merupakan salah satu algoritma pembelajaran mesin yang sering digunakan dalam klasifikasi dan regresi. Algoritma ini bekerja dengan mencari *hyperplane* atau batas pemisah terbaik antara kelas-kelas dalam data berdimensi tinggi. Menurut IBM, SVM digunakan untuk menemukan *hyperplane* optimal yang memaksimalkan margin antara kelas yang berbeda, sehingga data dari kelas yang berbeda terpisah secara jelas. Dalam konteks diagnosis depresi, SVM dapat memisahkan individu yang mengalami depresi dengan yang tidak, berdasarkan berbagai fitur seperti pola tidur, tingkat aktivitas fisik, dan kondisi medis

Algoritma SVM dapat menangani dataset dengan jumlah fitur yang besar, yang membuatnya ideal untuk dataset yang kompleks seperti yang digunakan dalam penelitian ini. SVM bekerja dengan menggunakan fungsi *kernel* untuk memetakan data ke ruang dimensi yang lebih tinggi, sehingga memungkinkan pemisahan yang lebih baik bahkan ketika data tidak dapat dipisahkan secara linear. Beberapa *kernel* yang sering digunakan antara lain *Linear Kernel*, *Polynomial Kernel*, dan *Radial Basis Function* (RBF).

SVM dapat menangani dataset dengan banyak fitur, yang membuatnya ideal untuk dataset yang kompleks seperti yang digunakan dalam penelitian ini. Untuk kasus data yang tidak dapat dipisahkan secara linear, SVM menggunakan fungsi *kernel* untuk memetakan data ke ruang dimensi yang lebih tinggi, sehingga memungkinkan pemisahan yang lebih baik. Beberapa *kernel* yang umum digunakan meliputi: *Linear Kernel*, *Polynomial Kernel*, *Radial Basis Function* (RBF)

Dalam penelitian kesehatan mental, SVM sering digunakan. Sacchet et al. (2015) [11] menggunakan SVM untuk mengklasifikasi gangguan depresi mayor berdasarkan data neuroimaging, dan algoritma ini berhasil memberikan akurasi yang tinggi dalam memisahkan kelas depresi mayor. Selain itu, Su et al. (2013) [12] juga menunjukkan keberhasilan SVM dalam mendeteksi depresi dari data fMRI dengan akurasi yang baik.

Visualisasi dari SVM biasanya menampilkan data di dua kelas dengan *hyperplane* yang memisahkan kedua kelas tersebut. Berikut adalah rumus dari *hyperplane* dengan *linear kernel* yang digunakan dalam SVM:

$$w \cdot x + b = 0,$$

dengan  $w$  adalah vektor bobot,  $x$  adalah vektor fitur, dan  $b$  adalah bias atau offset.

*Hyperplane* ini bekerja dengan memaksimalkan margin antara dua kelas, yaitu jarak terdekat dari titik data di masing-masing kelas ke *hyperplane*. Untuk kasus data yang tidak dapat dipisahkan secara linear, SVM menggunakan fungsi *kernel* seperti **Radial Basis Function (RBF)** untuk memetakan data ke ruang dimensi yang lebih tinggi, sehingga memungkinkan data terpisah dengan lebih baik.

#### B. Random Forest (RF)

Definisi dasar *Random Forest* menurut Turing.com [13], yaitu adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi, yang mana prediksi depresi peneliti dikategorikan sebagaimana. Algoritmanya juga membutuhkan banyak data, dan dataset peneliti menyediakan cukup; 196,851 nilai agar algoritma ini berguna.

Algoritmanya bekerja dengan membangun banyak pohon keputusan dengan 'Random Sampling with Replacement', yaitu mengambil fitur dan nilainya dari datasetnya secara acak, sampai '**Max Depth**' yaitu **Kedalaman Maksimum**, yang berarti beberapa panjang pohnnya. Kedalaman Maksimum itu ditentu oleh peneliti.

Karena sifat dataset peneliti dan apa yang ingin peneliti capa untuk penelitian ini, peneliti harus menggunakan klasifikasi, dan klasifikasi

menggunakan Random Forest itu sering dilakukan melalui Gini Index dengan rumus berikut:

$$Gini = 1 - \sum_{i=1}^C (p_i)^2$$

Alternatif selain Gini Index adalah menggunakan Entropy dengan rumus berikut:

$$Entropy = \sum_{i=1}^C p_i * \log_2(p_i)$$

Kedua rumus[14] diatas digunakan untuk menentukan bagaimana node-node akan bercabang dalam pohon keputusan.

Untuk mempermudah pemahaman, peneliti akan menggunakan penjelasan dan contoh dalam konteks dataset yang berkaitan dengan depresi manusia.

Langkah pertama adalah menentukan fitur-fitur yang relevan untuk algoritma. Dalam hal ini, fitur **Nama** akan dikecualikan dari model karena nama individu tidak memiliki relevansi dalam mempelajari depresi pada populasi secara keseluruhan. Pengecualian ini penting untuk memastikan bahwa model tidak terhambat oleh data yang tidak relevan, yang dapat mempengaruhi kinerja algoritma.

Setelah itu, proses pembuatan pohon keputusan akan dimulai. Contohnya, kita bisa memulai dengan fitur **Usia** sebagai node pertama. Node berikutnya mungkin didasarkan pada **Jumlah Anak**. Setelah itu, node selanjutnya bisa merujuk pada **Pola Tidur**. Proses ini akan berlanjut hingga mencapai **Kedalaman Maksimum** yang telah ditentukan. Ini hanya merupakan contoh untuk satu pohon keputusan. Pohon-pohon lainnya dapat memiliki fitur dan nilai yang berbeda-beda, menciptakan variasi dalam struktur dan hasil model.

Setelah itu, proses prediksi akan dilakukan dengan menggunakan metode **Voting**, di mana algoritma akan menentukan mayoritas untuk setiap nilai berdasarkan status depresi yang dinyatakan sebagai “Ya” atau “Tidak”. Sebagai contoh, individu yang berusia antara 30-35 tahun, memiliki pola tidur yang tidak cukup, dan mengkonsumsi alkohol, dapat dianggap mengalami depresi jika status depresi mereka lebih sering bernilai “Ya”

dibandingkan “Tidak”. Dengan pendekatan ini, algoritma akan mengidentifikasi pola-pola yang menunjukkan kecenderungan depresi berdasarkan kombinasi fitur yang dianalisis.

Para peneliti dari Tiongkok sudah menggunakan *Random Forest Classifier* untuk mendeteksi dan klasifikasi depresi untuk lansia cacat di pedesaan dan perkotaan[15].

### C. Extreme Gradient Boosting (XGBoost) Classifier

*XGBoost Classifier* merupakan penerapan dari pohon keputusan (*decision tree*) yang didorong oleh metode *gradient descent* dalam kerangka algoritma pembelajaran mesin *ensemble* [16]. Yang membedakan XGBoost dari algoritma pohon lainnya adalah penggunaan teknik *gradient boosting*, di mana terdapat penambahan model baru untuk memperbaiki kesalahan yang terjadi pada model sebelumnya. Pohon keputusan sendiri berfungsi untuk membagi dataset menjadi subset berdasarkan nilai fitur, sehingga *leaf nodes* atau ujung cabang pohon tersebut menghasilkan prediksi kategorinya. Algoritma yang dilatih ini direpresentasikan sebagai sekumpulan kondisi aturan *if and else* [17]. Dalam sebuah penelitian Sharma dkk. telah berhasil menerapkan XGBoost untuk diagnosis depresi. Mereka menggunakan teknik *oversampling* dan *over-under sampling* untuk mengatasi ketidakseimbangan data, dan berhasil mencapai *balanced accuracy* dan F1 score lebih dari 90% [18].

Ketika XGBoost diaplikasikan untuk melakukan klasifikasi terhadap dataset, model ini menggunakan dataset  $x$  untuk memprediksi hasil berupa  $\hat{Y}$ , yang didapat dari:

$$\hat{Y} = \sum_{i=1}^I f_i(x_i), f_i \in A,$$

dengan  $I$  adalah jumlah pohon keputusan,  $f_i$  adalah fungsi dalam ruang fungsi  $A$  untuk pohon ke- $i$ , dan  $A$  adalah himpunan dari semua pohon klasifikasi dan regresi yang mungkin.

Dalam pelatihan model, setiap pohon keputusan yang sudah dilatih akan mencoba untuk melengkapi *residual* melalui mengoptimalkan *fungsi D* pada pohon ke-(n+1) sebagai berikut:

$$D = \sum_{k=1}^n l(y_k, \hat{y}_k^{(n)}) + \sum_{k=1}^n \Omega(f_k),$$

dengan fungsi  $l$  adalah fungsi *loss* atau kesalahan prediksi,  $y_k$  adalah nilai output sebenarnya,  $\hat{y}_k^{(n)}$  adalah nilai prediksi pada langkah waktu ke-n. Fungsi  $\Omega$  adalah regularisasi untuk menghindari *overfitting* yang menghambat performa atau akurasi model ketika terdapat data yang baru, dengan cara menambahkan komponen tambahan pada fungsi loss sebagai *penalty term* yang didasarkan pada nilai parameter modelnya, sehingga fungsi loss dapat diminimalisir secara keseluruhannya.

Fungsi  $\Omega$  atau fungsi regularisasi dihitung sebagai berikut:

$$\Omega(f) = \gamma L + \frac{1}{2} \lambda \sum_{i=1}^L w_i^2,$$

dengan  $L$  adalah jumlah *leaf* atau ujung cabang dari pohon keputusan dan  $w_i$  adalah bobot atau nilai dari *leaf* ke- $i$ .

Algoritma XGBoost secara otomatis memilih fitur yang paling berpengaruh selama pelatihan untuk dijadikan *node* di dalam pohon sebagai pemisah antar kategori data. Pengaruh dari fitur dapat dihitung melalui pengukuran *information gain*, *gain ratio*, dan Gini index. Seperti *decision tree*, fitur yang memiliki nilai *information gain* yang terbaik akan dipakai terlebih dahulu sebagai data pemisah, karena *information gain* mengukur seberapa baik fitur tersebut memisahkan data yang digunakan untuk melatih algoritma sesuai dengan target klasifikasi. *Information gain* diperoleh dari perhitungan nilai *entropy* dari dataset, nilai *entropy* untuk setiap fitur, dan selisih antara *entropy* secara keseluruhan dengan perkalian dari kemungkinan dan *entropy* dari masing-masing *instance*, dengan rumus sebagai berikut:

$$Entropy(S) \equiv -p_1 \log_2 p_1 - p_2 \log_2 p_2$$

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

## IV. METODOLOGI

### A. Overview

#### 1) Support Vector Machine (SVM)

Penelitian dari Li Yi, et al. bertujuan untuk meningkatkan diagnosis depresi secara objektif dengan memanfaatkan kombinasi elektroensefalografi (EEG) dan spektroskopi inframerah-dekat (fNIRS). Data diperoleh dari 25 pasien dengan gangguan depresi mayor (MDD) dan 30 individu kontrol sehat. Pengambilan data dilakukan selama kondisi istirahat dengan mata tertutup, menggunakan perangkat EEG 32 saluran dan fNIRS delapan saluran. Sinyal EEG diproses untuk menghilangkan noise melalui *filter* pita 0,5–30 Hz, sementara sinyal fNIRS di-*filter* untuk menghilangkan noise frekuensi rendah dan drift.

Setelah data diproses, fitur utama seperti efisiensi lokal jaringan pada pita delta, asimetri hemisfer pada pita theta, dan entropi oksigenasi otak diekstraksi. Pemilihan fitur dilakukan dengan metode LASSO, yang secara efektif mengurangi dimensi data. Model klasifikasi dibangun menggunakan *Support Vector Machine* (SVM) dengan kernel linier, dan parameter model dioptimasi melalui validasi silang Leave-One-Out (LOOCV). Hasilnya menunjukkan akurasi 81,8% hanya dengan fitur EEG, yang meningkat menjadi 92,7% ketika fitur EEG dikombinasikan dengan fNIRS. Dengan demikian, penemuan ini menyoroti keunggulan pendekatan *hybrid* EEG dan fNIRS dalam mendeteksi depresi, menawarkan metode yang lebih objektif, non-invasif, dan berpotensi diterapkan secara luas dalam diagnosis klinis. [25]

#### 2) Random Forest (RF)

Peneliti Yu Xin dan Xiaohui Ren (2022) telah melakukan penelitian untuk memprediksi depresi di Tiongkok pada lansia pedesaan dan perkotaan menggunakan *Random Forest Classifier*. Ukuran sampelnya adalah 1460 orang, yang 841 diantaranya berasal dari daerah pedesaan dan 619 dari daerah perkotaan. 70% datanya digunakan untuk pelatihan dan 30% untuk pengujian.

Parameter berikut digunakan untuk daerah pedesaan:

$$\begin{cases} \text{'n\_estimators'} & : 128, \\ \text{'min\_samples\_split'} & : 16, \end{cases}$$

```

'min_samples_leaf'      : 8,
'max_depth'             : 7,
}.

```

Berikutnya untuk daerah perkotaan:

```

{'n_estimators'         : 190,
'min_samples_split'    : 16,
'min_samples_leaf'     : 8,
'max_depth'             : 7,
}.

```

Penelitian menyimpulkan bahwa tingkat depresi pada lansia pedesaan adalah 57,67% dan 44,59% untuk lansia perkotaan.

Prediktor yang paling mempengaruhi dan signifikan adalah:

- *Self Rated Health*
- *Changing in Perceived Health,*
- *Disease or Accidental Experience Within The Past 2 Weeks*
- *Life Satisfaction.*
- *Trusting People*
- *BMI*
- *Having Trust in the Future.*

Kinerja klasifikasi ini diuji menggunakan metode validasi silang 10-k dan menghasilkan akurasi sebesar 70% untuk pedesaan dan 71% untuk perkotaan.

### 3) Extreme Gradient Boosting (XGBoost)

Studi yang dilakukan oleh Sharma, Amita, dkk. (2020) telah menerapkan *Machine Learning*, khususnya dengan algoritma *XGBoost* untuk mendeteksi kasus depresi dari 11,081 observasi/warga dari Belanda. Karena diantaranya terdapat 570 kasus depresi, penelitian tersebut membandingkan model dari dataset pelatihan asli dan model dari data pelatihan melalui beberapa strategi *resampling* (pengambilan sampel ulang) guna mengatasi data yang tidak seimbang, yaitu *over-sampling*, *under-sampling*, *over-under sampling* dan *Random Over-Sampling Examples*. Berdasarkan hasil pelatihan model sebesar 80% dari data pelatihan, akurasi yang terbaik adalah model yang berasal dari data yang telah dilakukan *over-sampling* dengan *balanced accuracy* sebesar 97.65%, *F1 Score* sebesar 97.62%, akurasi sebesar 97.29%, presisi sebesar 95.48% dan *recall* sebesar 99.87%. Parameter yang ditetapkan adalah:

```

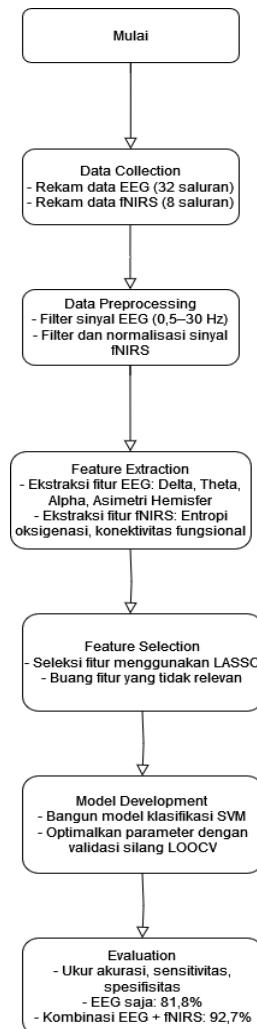
{'booster'              : "gbtree",
'objective'            : "binary: logistic",
'eta'                  : 0.3,
'gamma'                : 0,
'max_depth'             : 6,
'min_child_weight'     : 1,
'Subsample'             : 1,
'colsample_bytree'      : 1
}.

```

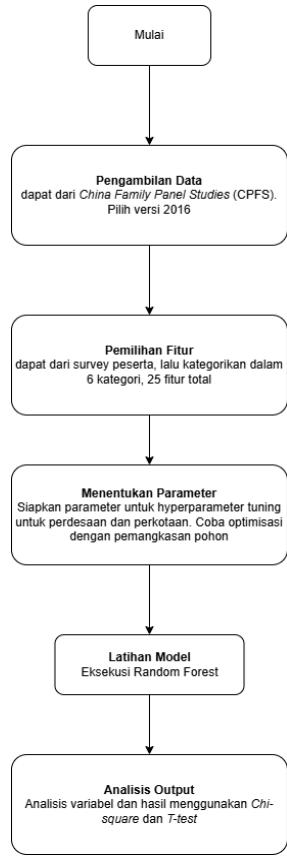
Berdasarkan *information gain*, beberapa fitur atau faktor terbesar yang berpengaruh terhadap status depresi secara berurut dari yang terbesar adalah *Neutrophil Granulocytes* (GR), *Uric Acid* (UZ), *Phosphate* (FOS), *Triglycerides* (TGL), *Cholesterol* (CHO), *Ureum* (UR), *Monocytes* (MO), *Thrombocytes* (TR), *Creatinine* (BKR), dan *Lymphocytes* (LY).

## B. Langkah/Arsitektur dan Tujuan dari Studi Lain

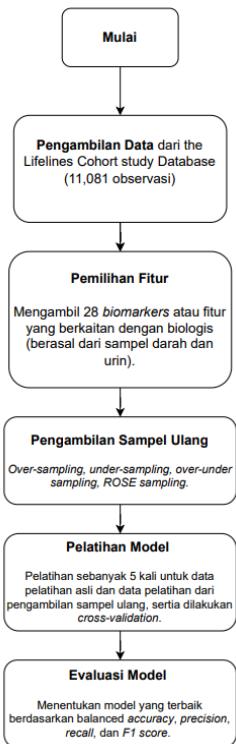
### 1) Support Vector Machine (SVM) dari Studi oleh Li Yi, et al.



- 2) Random Forest (RF) dari Studi oleh Yu Xin dan Xiaohui Ren



- 3) Extreme Gradient Boosting (XGBoost) dari Studi oleh Sharma, Amita, dkk.



### C. Input dan Output dari Studi Lain (Data Preprocessing dan Model)

- 1) Support Vector Machine (SVM) dari Studi oleh Li Yi, et al.

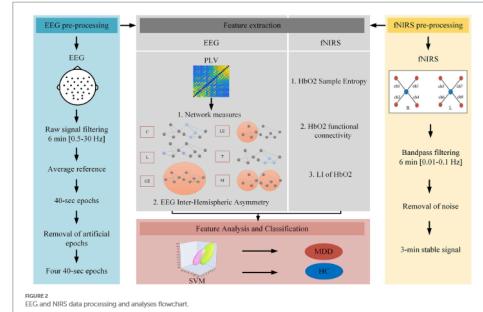


Diagram alur pemrosesan dan analisis data EEG dan fNIRS

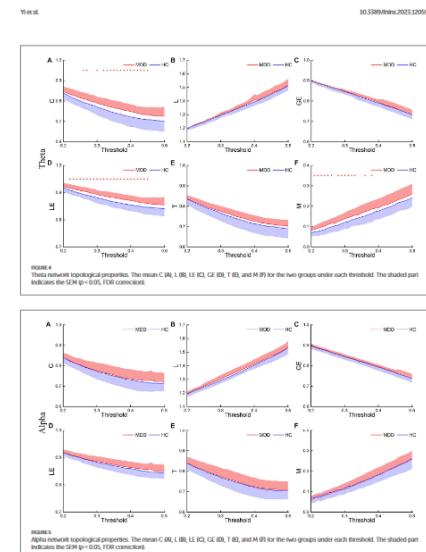


FIGURE 4 | Theta network topological properties. The mean C (A), L (B), LI (C), GE (D), T (E), and M (F) for the two groups under each threshold. The shaded part indicates the SEM ( $p < 0.05$ , FDR correction).



FIGURE 5 | Alpha network topological properties. The mean C (A), L (B), LI (C), GE (D), T (E), and M (F) for the two groups under each threshold. The shaded part indicates the SEM ( $p < 0.05$ , FDR correction).

Properti topologi jaringan Theta (atas) dan Alpha (bawah).

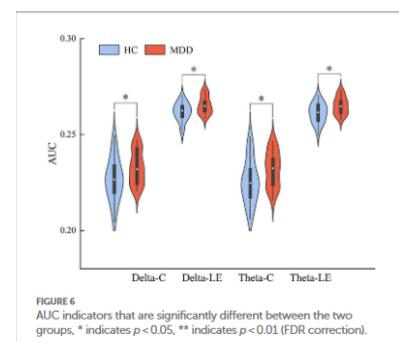


FIGURE 6 | AUC indicators that are significantly different between the two groups. \* indicates  $p < 0.05$ , \*\* indicates  $p < 0.01$  (FDR correction).

Indikator AUC yang berbeda signifikan antara dua kelompok.

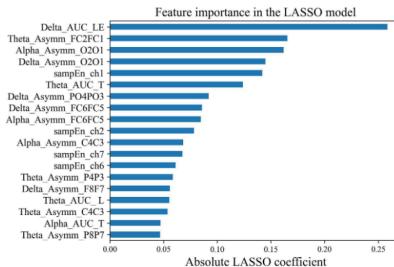


FIGURE 7  
Feature importance in the (EEG and fNIRS) LASSO Model.

### Feature Importance dalam Model LASSO

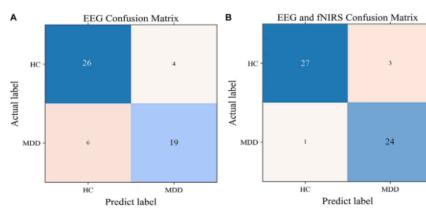


FIGURE 8  
The confusion matrix for the SVM model with EEG features (A) and with EEG and fNIRS features (B).

Confusion matrix untuk model SVM dengan fitur EEG (A) dan dengan fitur EEG dan fNIRS (B).

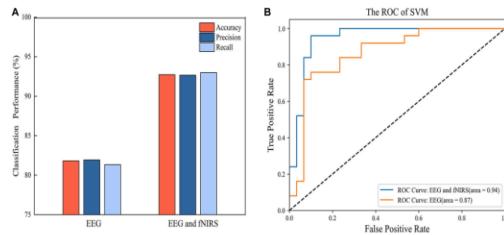


FIGURE 9  
(A) The classification performance of SVM with EEG features and hybrid EEG and fNIRS features. (B) The ROC curve of the SVM models.

Classification performance dan ROC curve model SVM.

## 2) Random Forest (RF) dari Studi oleh Yu Xin dan Xiaohui Ren

Table 2 Sociodemographic data and characteristics of rural and urban disabled elderly

Variables	Rural			Urban			P value*	
	Non-depression	Depression	P value	Sum	Non-depression	Depression		
<b>Demographical factors</b>								
Gender							0.636	
Male	164(46.07)	187(38.56)	0.029	351(41.74)	153(44.61)	113(40.94)	0.360	266(42.97)
Female	192(53.93)	298(61.44)		496(58.26)	190(55.39)	163(59.06)		353(57.03)
Age	71.35(0.38)	70.63(0.31)	0.019	70.93(0.24)	72.51(0.43)	72.44(0.47)	0.011	72.48(0.32)
Marital status							0.179	
Married	262(73.60)	311(64.12)	0.004	573(68.13)	251(73.18)	191(69.20)	0.277	442(71.41)
Unmarried/divorced/widowed	94(26.40)	174(35.88)		268(31.87)	92(26.82)	85(30.80)		177(28.59)
Years of education							0.000	
0	261(73.31)	384(79.18)	0.001	654(76.69)	218(63.56)	175(63.41)	0.092	393(63.49)
1~12	61(17.13)	43(8.87)		104(12.37)	90(26.24)	72(26.09)		162(26.17)
>12	34(9.55)	58(11.96)		92(10.92)	35(10.20)	29(10.51)		64(10.34)
Ln(family per capita income)	8.83(0.05)	8.67(0.04)	0.0218	13,271.75(2462.48)	9.63(0.06)	9.39(0.06)	0.004	24,088.76(1953.38)
<b>Health behavior</b>								
Smoking							0.352	
Yes	86(24.16)	98(20.21)	0.171	184(21.88)	71(20.70)	52(18.84)	0.564	123(19.87)
No	270(75.84)	387(79.79)		657(78.12)	272(79.30)	224(81.16)		496(80.13)
Drinking more than three times a week							0.828	
Yes	39(10.46)	37(7.69)	0.096	760(94.04)	391(11.37)	196(8.88)	0.057	589(9.37)
No	317(89.04)	448(92.37)		765(90.96)	304(88.63)	257(90.91)		561(90.63)
Sleep duration							0.064	
<6h	340(55.55)	97(20.00)	0.000	131(15.58)	36(10.50)	8028(99.00)	0.000	11610(374)
6-8h	110(30.90)	131(27.01)		241(28.66)	111(23.36)	84(30.43)		1593(51.50)
>8	212(59.55)	257(52.99)		465(55.77)	196(57.14)	112(40.58)		308(49.76)
Regular exercise							0.000	
Yes	132(37.08)	149(30.72)	0.053	281(33.41)	167(48.69)	112(40.58)	0.044	279(45.07)
No	224(62.92)	336(69.28)		560(66.59)	176(51.31)	164(59.42)		340(54.93)
Health status							0.147	
Chronic disease								
Yes	94(26.40)	229(47.22)	0.000	323(38.41)	125(36.44)	13649(28.00)	0.001	261(42.16)
No	262(73.60)	256(52.78)		518(61.59)	218(63.55)	140(50.72)		358(57.84)
BMI	2153(0.23)	2129(0.21)	0.0424	2157(0.15)	23.04(0.23)	22.32(0.26)	0.037	22.73(0.17)
Disease or accident experience within the past 2 weeks							0.005	
Yes	147(41.29)	335(69.07)	0.000	482(57.31)	133(38.78)	17663(77.00)	0.000	30949(92)
No	209(58.71)	150(30.93)		359(42.69)	210(61.22)	10036(23)		31050(08)
Hospitalization within 1 year							0.005	
Yes	82(23.03)	176(36.29)	0.000	298(30.68)	114(33.24)	119(43.12)	0.012	233(37.64)
No	274(76.97)	309(63.71)		583(66.32)	229(66.76)	157(56.88)		386(62.36)

Table 2 (continued)

Variables	Rural			Urban			P value*	
	Non-depression	Depression	P value	Sum	Non-depression	Depression		
<b>Self-rated health</b>								
Poor	127(35.67)	316(85.15)	0.000	443(52.68)	108(31.49)	170(61.59)	0.000	27844(91)
Fair	84(23.69)	83(16.91)		166(19.74)	92(26.82)	662(33.91)		15825(53)
Good	90(25.29)	68(14.02)		158(18.79)	97(28.28)	321(11.59)		12023(20.84)
Very good	36(10.11)	91(8.66)		45(5.33)	31(9.04)	62(17.17)		37(5.98)
Excellent	195(54)	102(9.09)		393(45)	154(37)	20(2.72)		172(2.5)
Changing in perceived health							0.006	
Better	23(6.46)	34(7.01)	0.000	57(6.78)	32(33)	103(6.2)	0.000	42(6.79)
Unchanged	133(37.36)	78(16.08)		211(25.09)	140(40.82)	62(22.46)		202(32.63)
Worse	200(56.18)	373(76.91)		573(68.13)	171(49.85)	204(73.91)		375(60.58)
<b>Family relations</b>								
Number of family members							0.000	
< 3	123(34.50)	18(17.32)	0.711	128(15.22)	123(35.86)	114(41.30)	0.383	133(21.49)
3-5	130(38.52)	170(35.05)		394(46.65)	153(44.31)	112(40.58)		332(53.63)
5+	103(28.93)	134(27.65)		319(37.93)	68(19.83)	50(18.12)		154(24.88)
Number of children							0.000	
<3	52(14.61)	76(15.67)	0.758	304(36.15)	71(20.70)	62(22.46)	0.714	237(38.29)
3-6	172(48.31)	222(45.77)		300(35.67)	189(55.10)	143(51.81)		264(42.65)
>6	132(37.08)	187(38.56)		237(28.18)	83(24.20)	71(25.72)		118(19.06)
Closing to children							0.921	
Yes	302(84.83)	398(82.06)	0.288	700(83.23)	299(87.17)	215(77.90)	0.002	514(83.04)
No	54(15.17)	87(17.94)		141(16.77)	44(12.83)	61(22.10)		105(16.96)
Receiving financial assistance from children							0.393	
Yes	197(55.34)	181(37.32)	0.032	501(59.57)	149(43.44)	126(45.65)	0.582	344(55.57)
No	159(44.66)	304(62.68)		340(40.43)	194(56.56)	150(54.35)		275(54.43)
Weekly family dinner							0.000	
Seven times	319(80.61)	405(82.51)	0.012	724(86.09)	301(87.76)	221(80.43)	0.012	523(84.49)
Less than seven times	37(10.39)	80(14.49)		117(13.91)	42(12.24)	54(19.57)		90(15.51)
<b>Social relations</b>								
Neighborhood help	17(70.05)	14(10.05)	0.000	158(0.04)	148(0.05)	16(0.06)	0.013	136(0.04)
Neighborhood relationship	22(50.38)	18(80.04)	0.000	21(0.03)	20(4.04)	22(1.05)	0.020	214(0.03)
Community emotion	2(01.04)	1(73.04)	0.000	1(89.03)	1(87.05)	2(18.06)	0.000	2(01.04)
Participating organizations	Yes	48(13.48)	61(12.58)	0.699	109(12.96)	100(29.15)	0.657	176(28.43)

Table 2 (continued)

Variables	Rural			Urban			P value*	
	Non-depression	Depression	P value	Sum	Non-depression	Depression		
<b>Trusting people</b>								
Yes	217(61.30)	258(53.75)	0.030	475(56.95)	206(60.41)	143(51.81)	0.032	349(56.50)
No	137(38.70)	222(46.25)		359(43.05)	135(39.59)	133(48.19)		268(43.44)
<b>Subjective attitude</b>								
Life satisfaction	4(90.05)	3(52.05)	0.000	3.76(0.04)	4.14(0.05)	3.55(0.07)	0.000	3.88(0.04)
Having trust in the future	3.88(0.06)	3.27(0.06)	0.000	3.53(0.04)	3.97(0.06)	3.36(0.07)	0.000	3.70(0.05)

## Data dan karakteristik sosiodemografi lansia cacat di pedesaan dan perkotaan

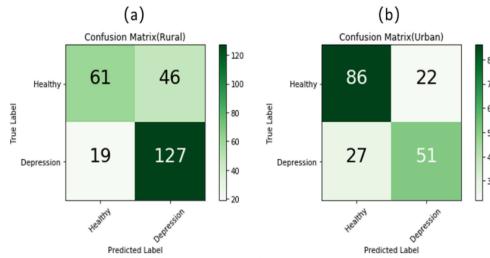
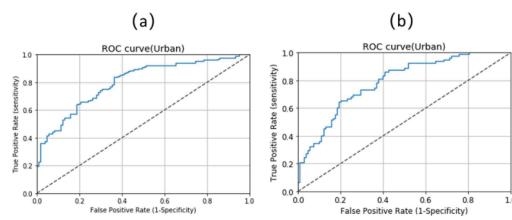


Fig. 2 a Confusion matrices for rural disabled elderly b Confusion matrices for urban disabled elderly

### Hasil Random Forest dalam bentuk Confusion Matrix



Kurva Receiver Operating Characteristic (ROC)

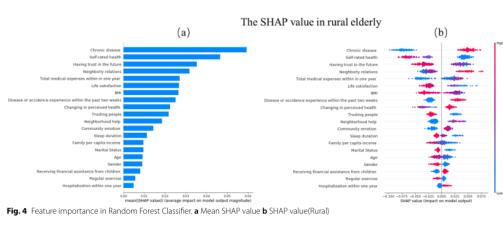
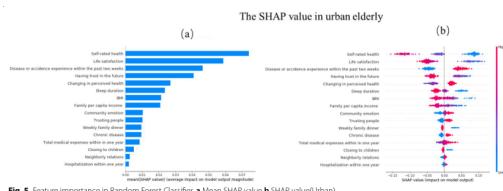


Fig. 4 Feature importance in Random Forest Classifier. a Mean SHAP value b Mean SHAP value(rural)



Tabel Shapley Additive Explanations (SHAP)

### 3) Extreme Gradient Boosting (XGBoost) dari Studi oleh Sharma, Amita, dkk.

Dataset yang disajikan dalam studi ini tidak tersedia untuk umum karena diperoleh dari Lifelines, yang menyediakan akses ke data melalui akses cloud yang aman setelah mendapatkan lisensi dan pembayaran biaya untuk jangka waktu terbatas. Untuk output yang dihasilkan sebagai berikut:

Confusion matrix elements	Original dataset (OR-Sample)	Under-sample dataset (U-sample)	Over-sample dataset (O-sample)	Over-under sample dataset (OU-sample)	ROSE sample dataset (R-sample)
Accuracy	0.9035	0.5164	0.9729	0.9442	0.6818
95% CI	0.8949, 0.9115	0.4724, 0.5603	0.9696, 0.9758	0.9378, 0.9502	0.6485, 0.6749
No Information rate	0.9488	0.5551	0.5838	0.5334	0.5107
p-value [Acc>NIR]	1.000	0.96495	<2.2e-16	<2.2e-16	<2.2e-16
Specificity	0.10166	0.5017	0.9982	0.9825	0.6699
Balanced accuracy	0.52413	0.5183	0.9765	0.9466	0.6617
Precision	0.0932	0.5017	0.9548	0.9107	0.6655
Recall	0.10	0.5737	0.9987	0.9835	0.6538
F1	0.0972	0.5183	0.9762	0.9457	0.6551

Source: Authors' own computation.

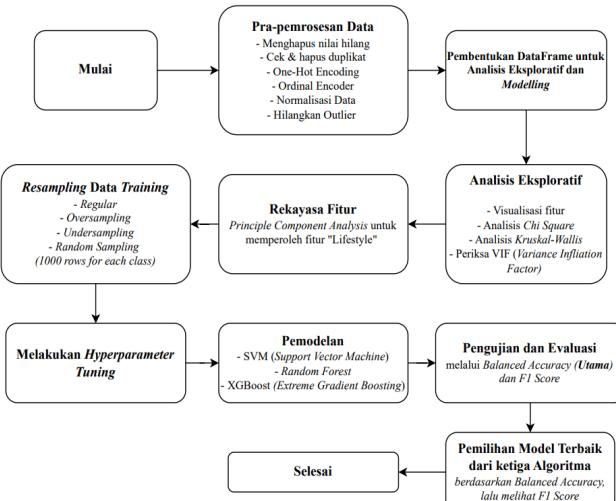
### Hasil Prediksi dari Pemodelan

Sr. no.	Xgb.OR	OR.Gain	Xgb.U	U.Gain	Xgb.O	O.Gain	Xgb.OU	OU.Gain	Xgb.R	R.Gain
1	GR	0.06	FOS	0.07	TR	0.06	TGL	0.07	GOT	0.11
2	UZ	0.08	LY	0.06	TGL	0.05	GR	0.06	AUB24	0.10
3	FOS	0.09	LDC	0.05	AF	0.05	LY	0.05	TGL	0.06
4	UZ	0.05	CA	0.05	UKR24	0.05	TR	0.05	GLU	0.05
5	CHO	0.05	BALB	0.05	ER	0.05	UKR24	0.05	AF	0.04
6	UR	0.05	TGL	0.05	ER	0.06	UR	0.05	HDC	0.04
7	MO	0.05	MO	0.05	LY	0.05	AF	0.05	LDC	0.04
8	TR	0.04	UZ	0.05	UR	0.05	FOS	0.04	FOS	0.03
9	BKR	0.04	TR	0.04	FOS	0.05	GLU	0.04	GR	0.03
10			CA	0.04	UZ	0.04	BA	0.04		
11	UKR24	0.04	AF	0.04	ALB24	0.04	HT	0.04	UKR24	0.03
12	CA	0.04	GR	0.04	GR	0.04	HT	0.04		
13	ER	0.04	CHD	0.04	GOT	0.04	ER	0.04	NA	0.03
14	EO	0.04	UKR24	0.03	GLU	0.04	ALB24	0.04	HB	0.03
15	ALB24	0.04	K	0.03	HT	0.03	CA	0.03	AST	0.03
16	AF	0.03	BKR	0.03	EO	0.03	ALT	0.03	MO	0.03
17	HB	0.03	ER	0.03	UZ	0.03	EO	0.03	BALB	0.03
18	HT	0.03	AST	0.03	GHO	0.03	CHO	0.03	BKR	0.03
19	DC	0.03	GLU	0.03	GR	0.03	HT	0.03	LY	0.03
20	GLU	0.03	AL	0.03	MO	0.03	LDC	0.03	UZ	0.03
21	AST	0.03	HT	0.03	BALB	0.03	BALB	0.03	AL	0.03
22	GOT	0.03	HDC	0.02	K	0.03	GOT	0.02	HT	0.03
23	ALT	0.02	ALB24	0.02	AST	0.03	MO	0.02	CA	0.03
24	K	0.02	HB	0.02	HDC	0.03	HDC	0.02	EO	0.03
25	HDC	0.02	BA	0.02	LDC	0.02	K	0.02	TR	0.02
26	BALB	0.02	GOT	0.01	HB	0.02	HB	0.02	CHO	0.02
27	BA	0.01	EO	0.01	NA	0.01	BA	0.02	K	0.02
28	NA	0.01	NA	0.01	BA	0.01	NA	0.01	ER	0.02

Source: Authors' own computation.

### Information Gain dari Fitur-fitur untuk Pemodelan

### D. Diagram Langkah/Arsitektur dari Penelitian Ini



- 1) **Pra-pemrosesan data** dilakukan dengan menghapus observasi yang tidak lengkap, data duplikat, serta melakukan transformasi pada data. Pada tahap transformasi, data numerik dinormalisasi, data ordinal diubah menjadi numerik berdasarkan peringkat (order) dan skalanya disesuaikan ke interval [0,1], sedangkan data nominal diubah menjadi representasi vektor biner. Selanjutnya, outlier dihapus dengan menggunakan batas  $z > 2$ .
- 2) Perbedaan antara *data frame* untuk pemodelan dan *data frame* untuk tahap EDA terletak pada transformasi data.

*Dataframe* untuk pemodelan sudah melalui tahap transformasi diatas. Sementara itu, *dataframe* untuk tahap EDA belum mengalami transformasi tersebut, sehingga data masih dalam bentuk aslinya untuk analisis eksploratif lebih lanjut.

- 3) **Analisis Eksploratif** dilakukan melalui **Uji Chi-Square** dan **Uji Kruskal Wallis** untuk menganalisis signifikansi hubungan antara setiap fitur dan variabel target.

Uji *Chi-Square* merupakan analisis statistika non-parametrik untuk mengukur hubungan antara dua variabel bertipe kategori (ordinal dan nominal) dengan membandingkan frekuensi yang diamati dan frekuensi yang diharapkan untuk menentukan apakah ada hubungan yang signifikan antara kedua variabel tersebut. Tujuan utamanya adalah untuk menguji apakah distribusi frekuensi pada dua variabel yang dikategorikan tidak saling berhubungan ( $H_0$ ), yang berarti distribusi frekuensi pada satu variabel tidak dipengaruhi oleh distribusi frekuensi pada variabel lainnya, atau terdapat hubungan antara kedua variabel tersebut ( $H_a$ ).

Berikut rumus dari uji *Chi-Square W*:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

df (derajat kebebasan) =  $(n-1)*(n-1)$ .

Dengan  $\chi^2$  adalah nilai statistik *Chi-Square*,  $O_i$  adalah frekuensi yang diamati pada sel ke-i dalam tabel kontingensi,  $E_i$  adalah frekuensi yang diharapkan pada sel ke-i jika tidak ada hubungan antara variabel, n adalah jumlah kelompok pada variabel yang diuji (yang diasumsikan memiliki jumlah kategori atau kelompok yang sama), dan  $\Sigma$  adalah operator untuk menjumlahkan hasil perhitungan untuk setiap sel dalam tabel kontingensi [19].

Selain itu, uji *Chi-Square* juga melibatkan tingkat signifikansi atau alpha ( $\alpha$ ), yang merupakan batas probabilitas yang

ditetapkan untuk menentukan apakah hasil uji signifikan atau tidak, dengan nilai umum 0.05 (5%). Jika  $p\text{-value} \leq \alpha$ , kita menolak hipotesis nol ( $H_0$ ), yang berarti ada hubungan signifikan antara dua variabel. Jika  $p\text{-value} > \alpha$ , kita gagal menolak  $H_0$ , yang menunjukkan tidak ada cukup bukti adanya hubungan. Dengan kata lain,  $p\text{-value}$  mengukur seberapa kuat bukti yang kita miliki untuk menolak hipotesis nol.

Uji *Kruskal-Wallis* merupakan analisis statistika non-parametrik untuk membandingkan dua atau lebih sampel yang independen dengan menganalisis populasi median. Jika tidak signifikan, maka tidak terdapat perbedaan antar sampel ( $H_0$ ). Namun, jika signifikan, maka terdapat perbedaan antar sampel ( $H_a$ ). Syarat data dari kedua sampel ini adalah berskala ordinal dan numerik. Perhitungan pengujian ini dimulai dari menggabungkan semua sampel dan mengurutkan nilai-nilai dari kedua sampel berdasarkan peringkat [20].

- 4) Setelah melakukan analisis uji signifikansi, peneliti melakukan **analisis multikolinearitas**. Analisis Multikolinearitas bertujuan untuk melakukan pemilihan fitur supaya mengurangi kesulitan algoritma atau model dalam membedakan antara fitur-fitur yang berkorelasi tinggi, sehingga salah satu fitur diantaranya harus dihilangkan. Analisis ini dilakukan dengan menguji **Korelasi Koefisien Pearson** antar data kontinu dan **Variance of Inflation (VIF)**.

Uji Korelasi Pearson digunakan untuk mengukur korelasi (hubungan) secara linier antara dua variabel (misalkan X dan Y) berskala kontinu. Koefisien korelasi atau  $r$  memiliki rentang nilai antara -1 dan 1. Nilai yang diantara mendekati 1 menunjukkan hubungan positif yang kuat antara kedua variabel; nilai yang diantara 0.4 dan 0.7 menunjukkan hubungan positif yang moderat; nilai yang mendekati 0 menunjukkan hubungan yang lemah atau

tidak ada hubungan; nilai yang diantara -0.4 dan -0.7 menunjukkan hubungan negatif yang moderat; nilai yang mendekati -1 menunjukkan hubungan negatif yang kuat antara kedua variabel.

Analisis *Variance Inflation Factor* (VIF) adalah metode uji multikolinearitas yang digunakan untuk mendeteksi apakah terdapat korelasi antar variabel independen dalam suatu model regresi, dengan mengidentifikasi kemungkinan adanya hubungan antar variabel bebas di dalam model regresi. Untuk menentukan apakah terjadi multikolinearitas dapat melalui nilai tolerance dan VIF. Pada umumnya, fitur yang memiliki multikolinearitas tinggi memiliki nilai VIF paling besar dan lebih besar daripada 5. Nilai toleransi mengukur proporsi variabilitas variabel independen yang tidak dapat dijelaskan oleh variabel bebas lainnya. VIF pada koefisien regresi untuk variabel-j dapat dihitung menggunakan rumus berikut:

$$VIF_j = \frac{1}{1-R_j^2},$$

dengan  $R_j^2$  merupakan koefisien determinasi antara variabel  $X_1, X_2, \dots, X_j$  dengan variabel bebas lainnya dalam model [21].

- 5) Proses **Rekayasa Fitur** (*Principal Component Analysis*) bertujuan untuk mendapatkan fitur "Gaya Hidup" yang komprehensif dari variabel-variabel tersebut: "Physics\_Activity", "Alcohol\_Consumption", "Diet\_Habits", "Sleep\_Patterns", "Current" (perokok aktif), dan "Former" (mantan perokok). Transformasi ini bertujuan untuk menciptakan keseimbangan yang lebih stabil antara data numerik dan kategorikal, meningkatkan kualitas dataset secara keseluruhan dan berpotensi meningkatkan akurasi model. Pada umumnya, *Principal Component Analysis* merupakan analisis korespondensi untuk menangani variabel kualitatif dan sebagai sebagai analisis faktor berganda untuk menangani kumpulan variabel yang heterogen.

Analisis ini populer untuk mengurangi dimensi set data numerik, memiliki ekstensi yang memungkinkannya digunakan untuk data kualitatif (kategori) dan untuk set data yang mengandung variabel heterogen (baik numerik maupun kategori) [22].

- 6) **Resampling data training** pada dataset Non-PCA dan PCA menggunakan *oversampling* (kecuali algoritma SVM karena *high computational cost*), *undersampling*, dan *random sampling* bertujuan untuk mengatasi ketidakseimbangan kelas. *Oversampling* menambah data kelas minoritas, *undersampling* mengurangi data kelas mayoritas, dan *random sampling* memilih subset acak (pada penelitian ini memilih 1000 dari setiap kategori) untuk menciptakan variasi dalam dataset. Metode ini digunakan untuk membuat 8 percobaan dan memilih model terbaik berdasarkan performa pada kedua dataset. Hal ini sudah termasuk perbandingan terhadap *dataset* yang tidak dilakukan proses ini (*dataset regular*).
- 7) **Hypertuning** bertujuan untuk mengoptimalkan *hyperparameter* model untuk meningkatkan kinerjanya dalam prediksi. Dalam *hypertuning*, berbagai kombinasi seperti *learning rate*, jumlah pohon, kedalaman pohon, dan parameter lainnya diuji menggunakan teknik seperti *grid search* atau *random search*. Kemudian, langkah ini menemukan kombinasi terbaik yang menghasilkan model dengan performa optimal berdasarkan metrik evaluasi tertentu, seperti *accuracy* dan *balanced accuracy*. Hal ini membantu model supaya lebih akurat dan tergeneralisir terhadap data yang tidak terlihat dalam *data training*.
- 8) **Pemodelan** dilakukan dengan menerapkan tiga algoritma *machine learning* yang telah dioptimalkan melalui *hypertuning*. Kemudian, **pengujian dan evaluasi model** dilakukan dengan membagi data menjadi *training* dan *testing set*, kemudian dievaluasi menggunakan nilai *balanced accuracy*

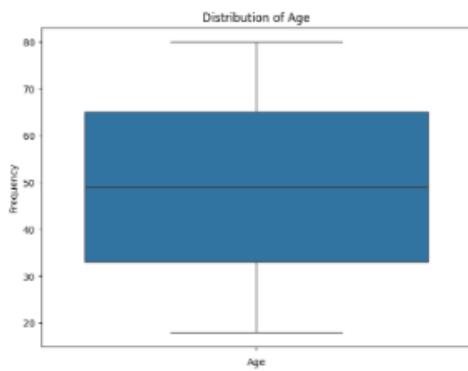
sebagai metrik utama. Jika akurasi antar model serupa, perbandingan dilakukan menggunakan *F1 Score* untuk memastikan keseimbangan antara *precision* dan *recall*. Selain itu, analisis *feature importance* digunakan untuk mengidentifikasi variabel yang paling berpengaruh terhadap hasil prediksi model. Untuk pemahaman lebih mendalam, SHAP (*SHapley Additive exPlanations*) values juga dianalisis untuk memberikan wawasan tentang kontribusi setiap fitur terhadap prediksi model dari kumpulan observasi. Terakhir, pemilihan akhir model ditentukan berdasarkan metrik utama dan metrik sekunder.

## V. HASIL, EVALUASI, DAN KESIMPULAN

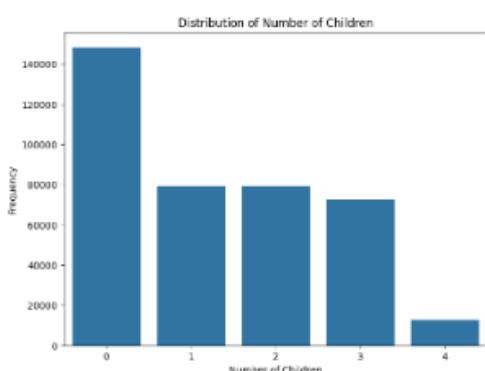
### A. Hasil

#### I) Analisis Eksploratif Data (EDA)

##### a) Variabel Bebas

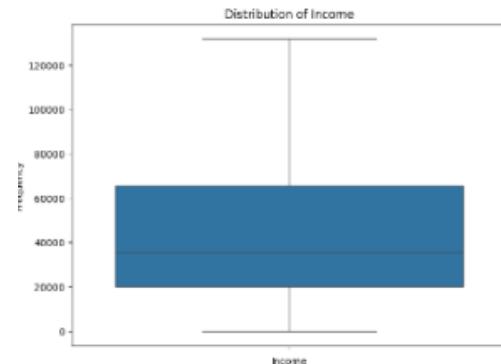


Data usia menunjukkan variasi dengan rentang yang cukup luas dan seimbang tanpa adanya outlier ekstrem. Sebagian besar data berasal dari responden berusia sekitar 30-65, dengan nilai minimum dan maksimum yang berbeda jauh.

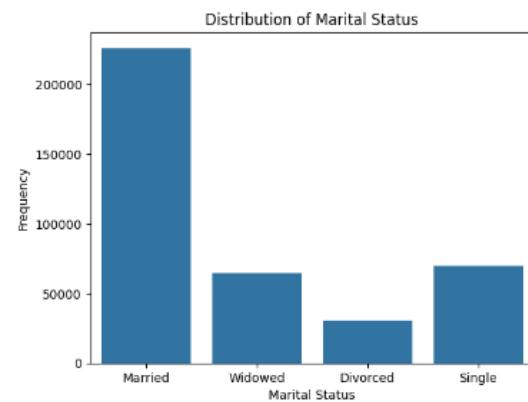


Data jumlah anak menunjukkan variasi responden dengan rentang antara 0-4 anak. Sebagian besar

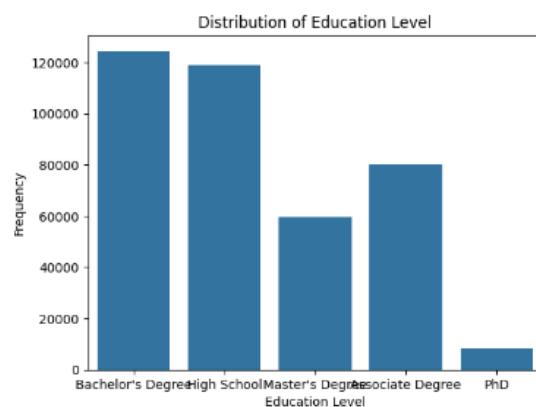
data berasal dari responden tidak memiliki anak, sedangkan jumlah anak antara 1 hingga 3 terdistribusi relatif secara merata.



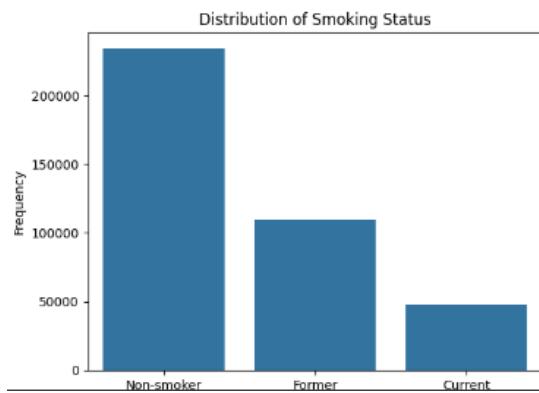
Data pemasukan atau gaji seseorang menunjukkan distribusi yang cenderung ke kiri karena memiliki nilai median yang relatif rendah (tidak di tengah *plot*). Sebagian besar data berasal dari responden yang memiliki gaji di antara \$20,000-\$60,000 per tahun.



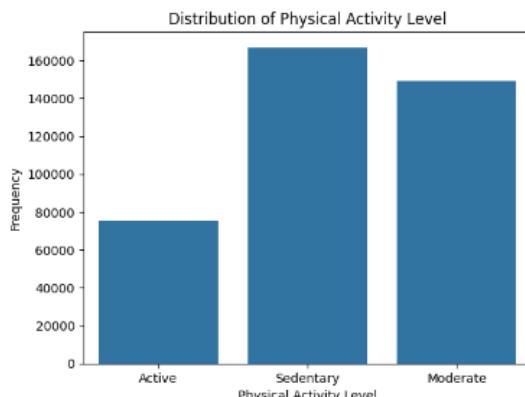
Data status pernikahan didominasi oleh responden dengan status sudah menikah, sedangkan sebagian kecil responden memiliki status cerai.



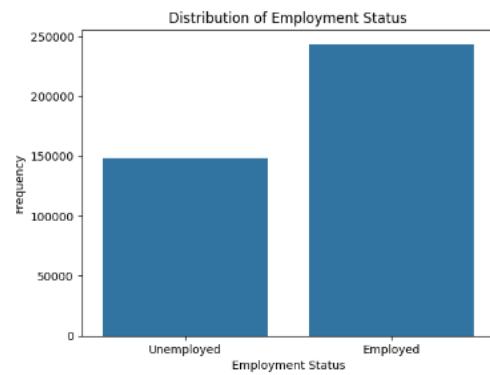
Data tingkat pendidikan menunjukkan bahwa dominasi responden menempuh gelar sarjana, SMA, dan *Associate Degree*, yang menandakan bahwa secara keseluruhan, distribusi responden ini berpendidikan menengah hingga tinggi. Sementara itu, frekuensi gelar PhD sangat rendah, yang menandakan bahwa sebagian kecil responden memiliki tingkat pendidikan tertinggi.



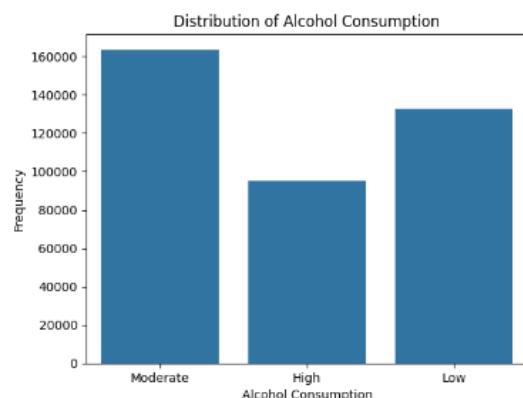
Data status merokok menunjukkan distribusi bahwa sebagian besar responden adalah bukan perokok, yang diikuti oleh mantan perokok dan perokok aktif.



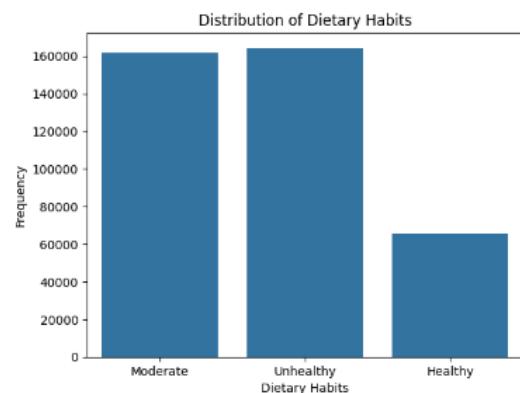
Data aktivitas fisik menunjukkan distribusi bahwa sebagian kecil responden aktif melakukan aktivitas fisik, sedangkan mayoritas responden jarang melakukan aktivitas fisik.



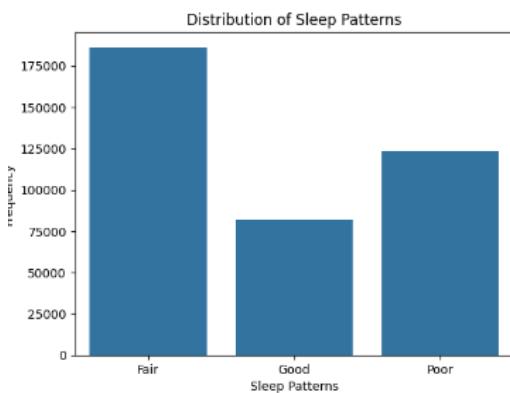
Data status kerja menunjukkan bahwa sekitar 60% responden sedang bekerja, sedangkan sisanya adalah responden yang tidak memiliki pekerjaan.



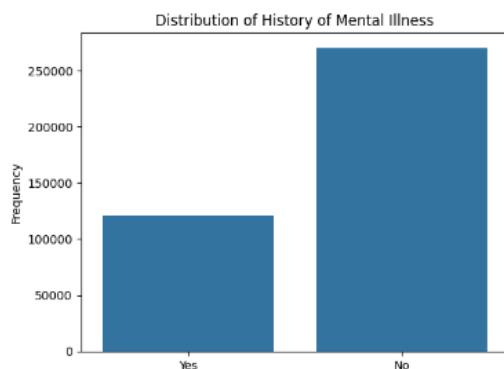
Data *alcohol consumption* menunjukkan bahwa sebagian besar responden mengkonsumsi alkohol dengan kadar sedang, diikuti oleh responden yang mengkonsumsi alkohol kadar rendah, dan kemudian responden yang mengkonsumsi alkohol kadar tinggi.



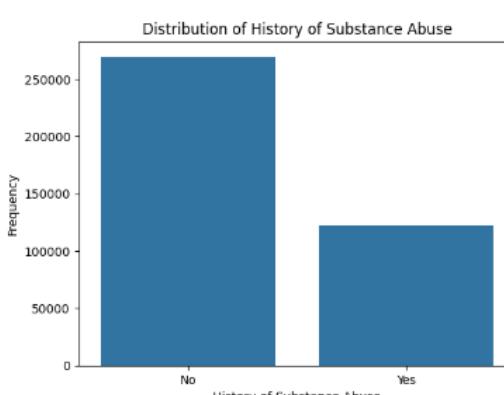
Data mengenai pola makan menunjukkan kecenderungan yang cukup signifikan di mana sebagian besar responden cenderung memiliki pola makan yang tidak sehat atau sedang. Responden dengan pola makan yang tidak sehat mendominasi, diikuti oleh responden dengan pola makan sedang, yaitu antara sehat dan tidak sehat.



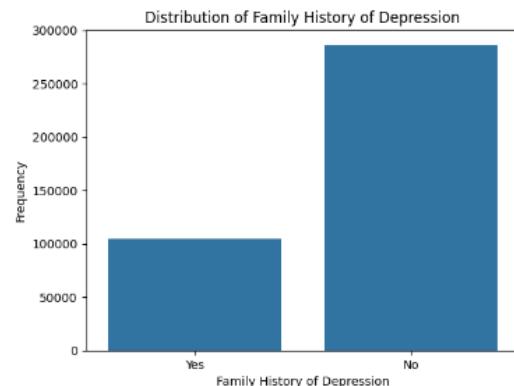
Data mengenai pola tidur menunjukkan kecenderungan yang cukup signifikan di mana sebagian besar responden cenderung memiliki kualitas pola tidur yang sedang. Responden dengan pola tidur yang sedang, diikuti oleh responden dengan pola tidur yang tidak baik, dan kemudian pola tidur yang sehat.



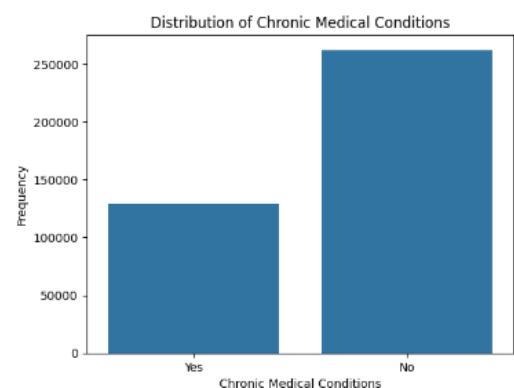
Data mengenai status depresi sebagai variabel target menunjukkan bahwa sebagian besar responden, yaitu sekitar 66% tidak terkena depresi atau gangguan mental, yang menunjukkan bahwa distribusi data untuk melakukan pemodelan algoritma pembelajaran mesin tidak seimbang, sehingga perlu dilakukan pengambilan sampel berulang atau *resampling*.



Data mengenai penggunaan obat terlarang menunjukkan bahwa sebagian besar responden, yaitu sekitar 70% tidak mengonsumsi obat terlarang.



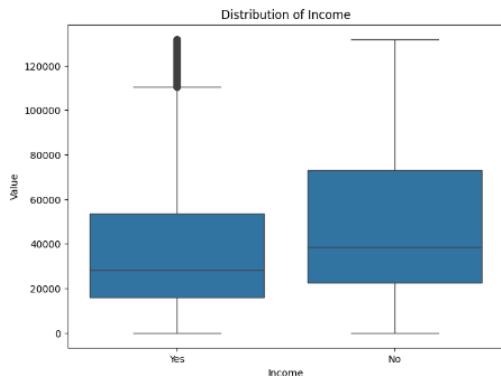
Data mengenai penggunaan obat terlarang menunjukkan bahwa sebagian besar responden, yaitu sekitar 70% tidak mengonsumsi obat terlarang.



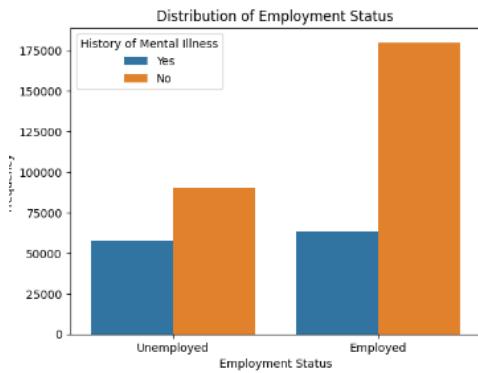
Data mengenai kondisi medis klinik menunjukkan bahwa sebagian besar, yaitu sekitar 66% responden tidak mengalami kondisi klinis secara media.

#### b) Variabel Bebas terhadap Variabel Target

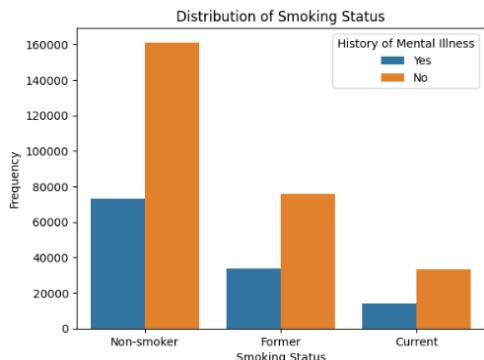
Data visualisasi bertujuan untuk mempelajari distribusi dari data berdasarkan perbedaan kelas atau status depresi dari para responden. Namun, perlu diperhatikan bahwa diperlukan uji signifikansi statistika untuk menganalisis hubungan antara setiap variabel bebas dengan variabel tergantung (status depresi). Walaupun terdapat banyak fitur yang dapat divisualisasikan, penelitian ini membahas hasil analisis dan visualisasi dari beberapa fitur yang memiliki *feature importance* yang lebih tinggi daripada fitur-fitur lainnya, diantaranya:



Secara visualisasi, berdasarkan status depresi, distribusi dari data pemasukan/gaji menunjukkan bahwa ada kecenderungan perbedaan median gaji antara responden yang depresi dengan yang tidak mengalami depresi, karena semakin besar.

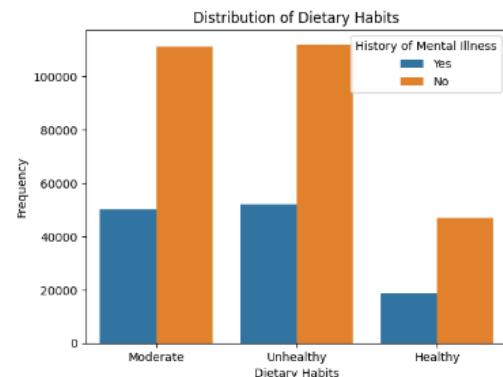


Secara visualisasi, berdasarkan status depresi, distribusi dari status kerja menunjukkan bahwa responden yang sedang tidak bekerja memiliki proporsi terkena depresi lebih tinggi daripada responden yang sedang bekerja.

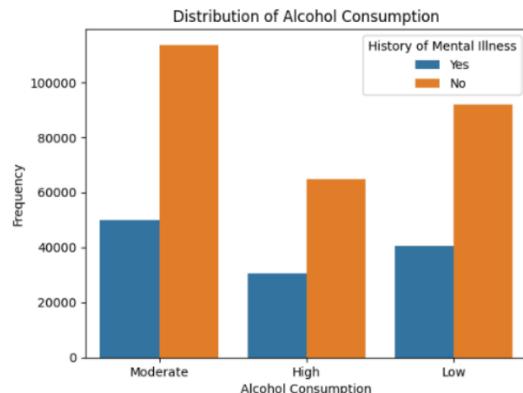


Secara visualisasi, berdasarkan status depresi, distribusi dari status merokok menunjukkan bahwa responden yang tidak merokok cenderung memiliki proporsi terkena depresi lebih rendah daripada status lainnya. Namun, responden yang sedang

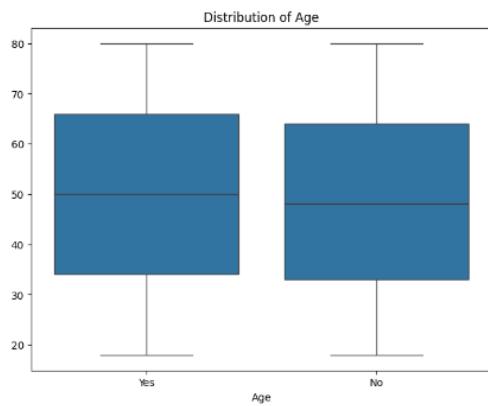
merokok cenderung memiliki proporsi terkena depresi terbesar.



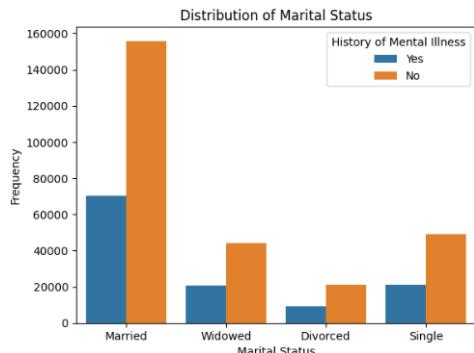
Secara visualisasi, berdasarkan status depresi, distribusi dari status merokok menunjukkan bahwa responden yang memiliki pola makan yang tidak sehat dan sedang cenderung memiliki proporsi terkena depresi lebih rendah daripada responden yang sedang menjalani pola makan yang sehat.



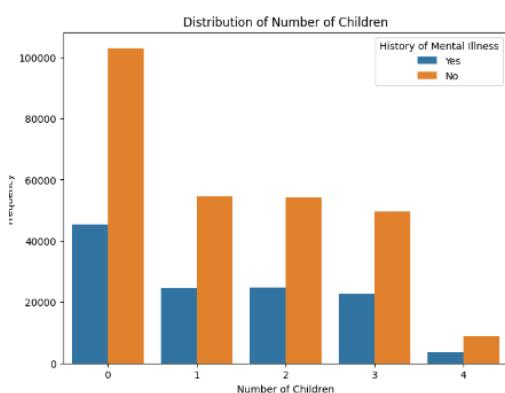
Secara visualisasi, berdasarkan status depresi, distribusi dari kadar konsumsi alkohol menunjukkan bahwa responden yang mengkonsumsi alkohol berkadar tinggi cenderung memiliki proporsi terkena resiko depresi terbesar, yaitu sekitar 30%, dibandingkan dengan kelompok-kelompok pengkonsumsi alkohol lainnya.



Secara visualisasi, berdasarkan status depresi, distribusi dari data usia menunjukkan bahwa tidak ada kecenderungan perbedaan median usia antara responden yang depresi dengan yang tidak mengalami depresi.



Secara visualisasi, berdasarkan status depresi, distribusi dari status pernikahan menunjukkan bahwa responden yang sedang menikah cenderung memiliki proporsi terkena depresi lebih tinggi daripada responden yang tidak menikah dan berstatus duda. Namun, responden yang telah bercerai cenderung memiliki proporsi terkena depresi terbesar, yaitu sekitar 50%.



Secara visualisasi, berdasarkan status depresi, distribusi dari data jumlah menunjukkan bahwa responden yang tidak memiliki anak cenderung memiliki proporsi status depresi yang lebih tinggi

daripada kelompok-kelompok responden lainnya yang memiliki 1-3 anak. Meskipun proporsi responden yang memiliki 1-3 anak dan mengalami depresi cenderung saling serupa, kelompok responden yang memiliki 4 anak memiliki proporsi status yang mengalami depresi paling tinggi.

### c) Analisis Hubungan Variabel Bebas terhadap Variabel Target

Berikut adalah hasil uji signifikansi statistika antara fitur-fitur terhadap variabel target:

No .	Fitur	p-value dari uji Chi-Square terhadap status depresi/gangguan mental
1	Marital Status	0.0
2	Education Level	0.7886
3	Smoking Status	4.7606e-07
4	Physical Activity Level	4.7606e-07
5	Employment Status	0.0
6	Alcohol Consumption	3.7468e-14
7	Dietary Habits	2.0122e-49
8	Sleep Patterns	8.4199e-24
9	History of Substance Abuse	0.3187
10	Family History of Depression	0.0019
11	Chronic Medical Conditions	0.0006

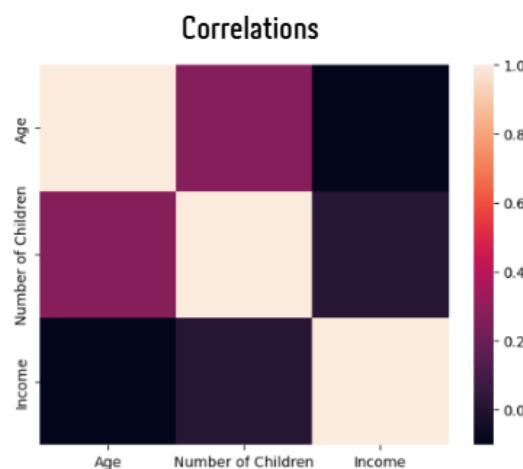
No .	Fitur	p-value dari Uji Kruskal-Wallis terhadap status depresi/gangguan mental
1	Age	1.1299e-68
2	Number of	0.0007

	Children	
3	Income	0.0

Berdasarkan hasil pengujian, belum ada bukti yang signifikan untuk membedakan fitur *Education Level* dan *History of Substance Abuse* terhadap status depresi (kedua *p-value* lebih besar daripada 0.05), sehingga kedua fitur tersebut harus dihilangkan untuk mengurangi *noise* dari *dataset*.

#### d) Analisis Multikolinearitas

Berikut adalah hasil analisis multikolinearitas setelah menghilangkan beberapa fitur setelah dilakukan *One Hot Encoding* dan *One Encoder* dengan VIF tertinggi dan lebih dari 5:



No	Fitur	VIF
1	Age	1.958022
2	Number of Children	1.413918
3	Income	2.236119
4	Divorced (Bercerai, 0 atau 1)	1.419570
5	Married (Menikah, 0 atau 1)	3.885681
6	Widowed (Duda, 0 atau 1)	2.562201
7	Current (Perokok Aktif, 0 atau 1)	1.377345
8	Former (Mantan)	1.433771

	Perokok, 0 atau 1)	
9	Physical Activity Level	2.565697
10	Alcohol Consumption	2.349300
11	Dietary Habits	2.456110
12	Sleep Patterns	2.709240
13	Employment Status	3.535282
14	Family History of Depression	1.430280
15	Chronic Medical Conditions	1.472557

Berdasarkan uji korelasi, tidak terdapat korelasi antara usia, jumlah anak, dan pemasukan/gaji. Sedangkan untuk uji VIF, fitur-fitur yang berada di dalam tabel akan digunakan untuk melakukan penelitian lebih lanjut.

#### 2) Rekayasa Fitur melalui Principal Component Analysis

Berikut adalah deskripsi data dari hasil pembuatan fitur baru “*Lifestyle*”

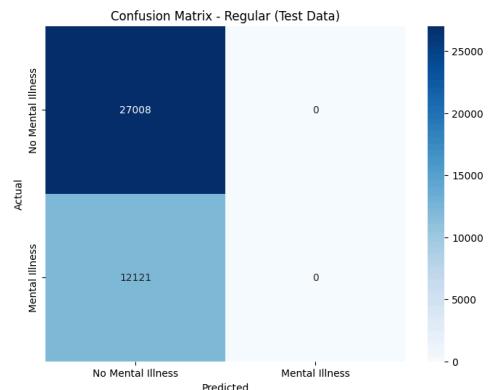
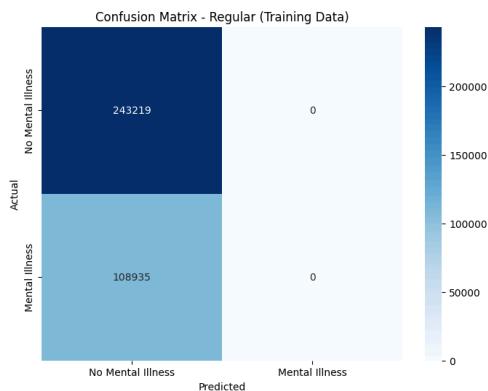
PCA	Lifestyle
count	391,283
mean	-0.0000000000000001854065
std	0.4635924
min	-0.7410305
25%	-0.2805144
50%	-0.1807609
75%	0.6109591
max	0.8429540

#### 3) Pemodelan a) Support Vector Machine (SVM)

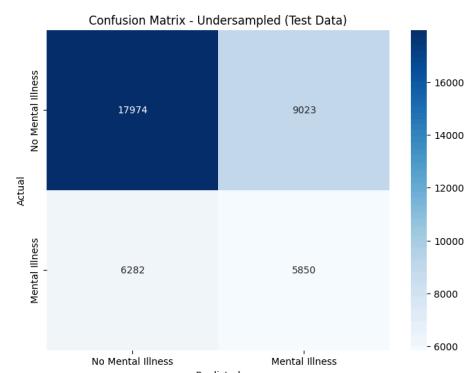
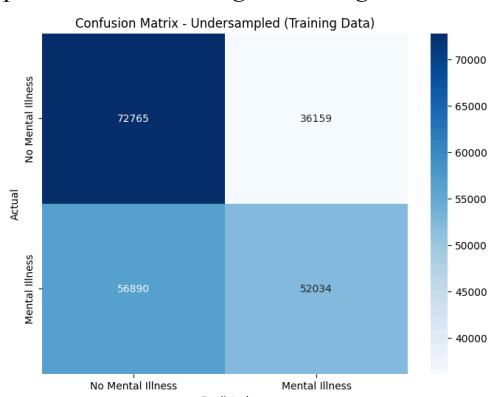
Untuk semua *trial*, best parameternya adalah

```
{C: [0.1, 1]
kernel: ['linear']
scoring: 'accuracy'
cv: 3
n_jobs: -1
random_state: 42}.
```

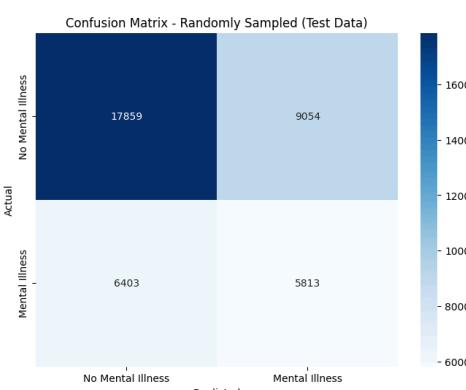
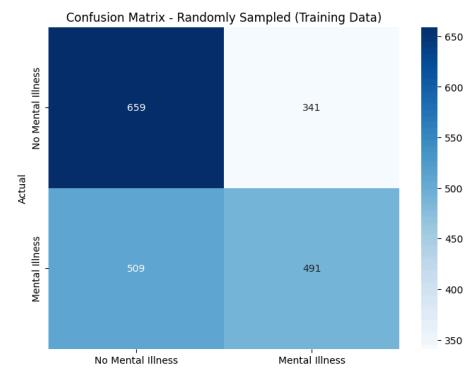
Berikut adalah hasil prediksi SVM *reguler* tanpa PCA untuk *training* dan *testing*:



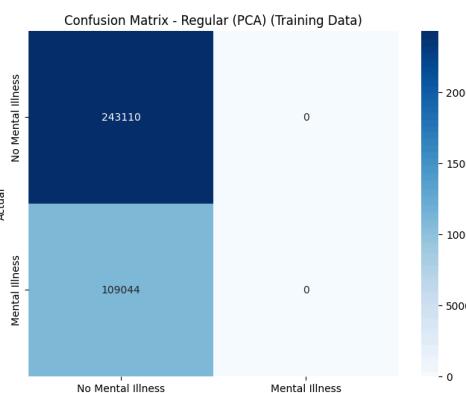
Berikut adalah hasil prediksi SVM *undersampled* tanpa PCA untuk *training* dan *testing*:

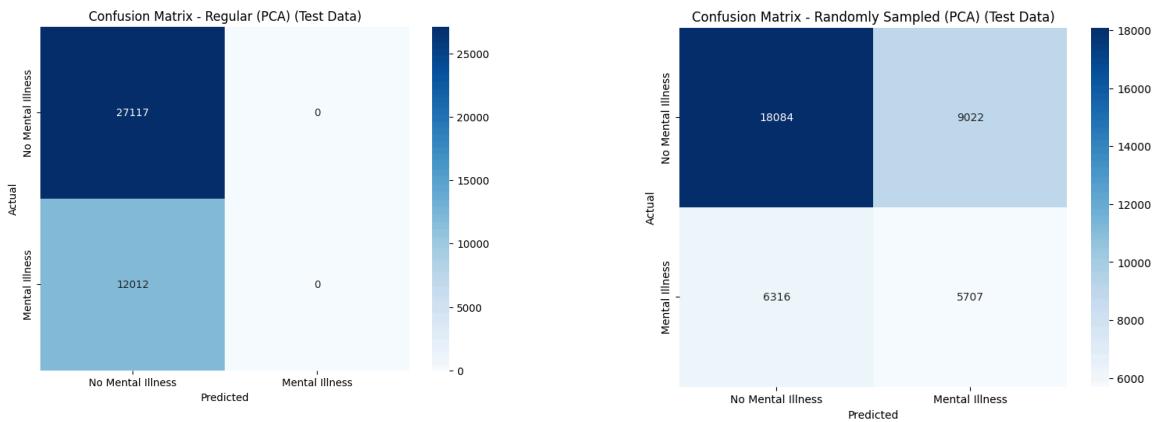


Berikut adalah hasil prediksi SVM *randomly sampled* tanpa PCA untuk *training* dan *testing*:

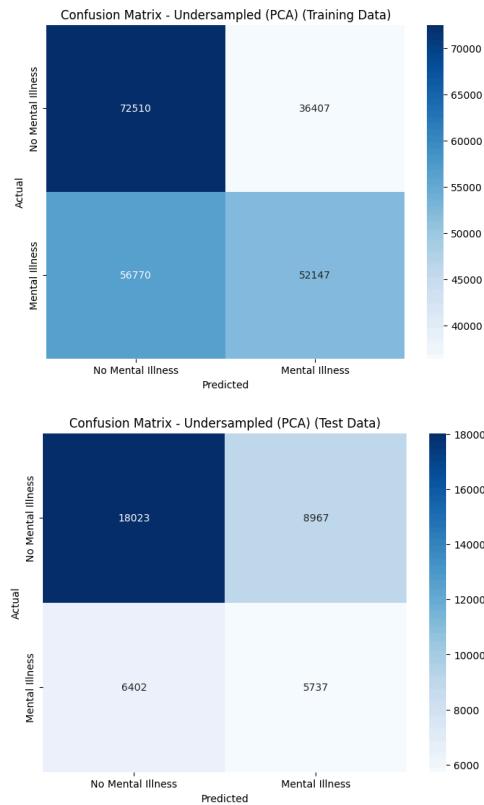


Berikut adalah hasil prediksi untuk SVM *regular* dengan PCA *training* dan *testing*:

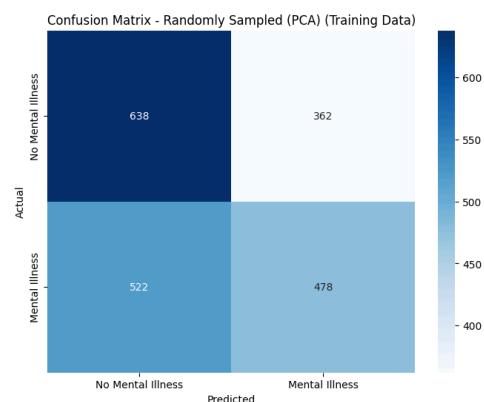




Berikut adalah hasil prediksi SVM *undersampled* dengan PCA untuk *training* dan *testing*:



Berikut adalah hasil prediksi SVM *randomly sampled* dengan PCA untuk *training* dan *testing*:

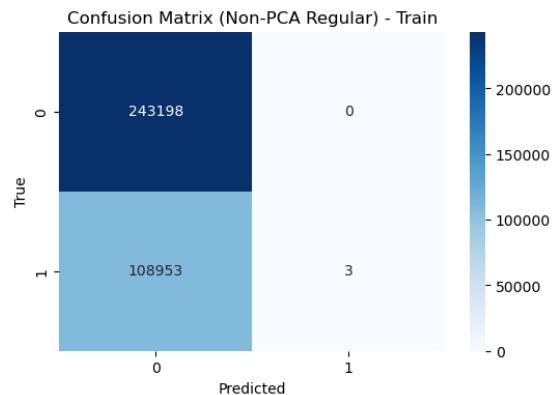


### b) Random Forest Classifier

Hasil *Random Forest regular* tanpa PCA dengan parameter terbaik yaitu:

```
{'n_estimators': 200,
'min_samples_split': 5,
'min_samples_leaf': 1,
'max_depth': 10,
}.
```

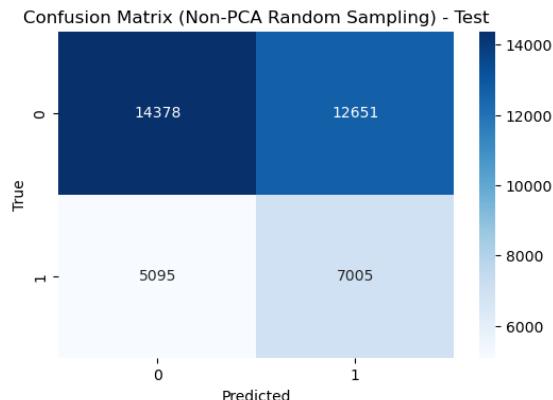
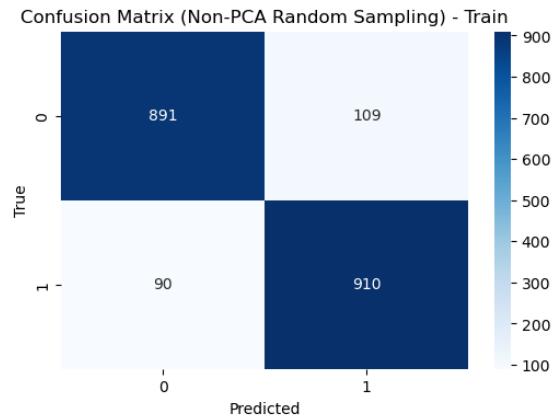
Berikut adalah hasil prediksi untuk *training* dan *testing*:



Hasil *Random Forest* yang dilakukan dengan *random sampling* tanpa PCA dengan parameter terbaik yaitu:

```
{'n_estimators': 200,
 'min_samples_split': 5,
 'min_samples_leaf': 1,
 'max_depth': 10,
 }.
```

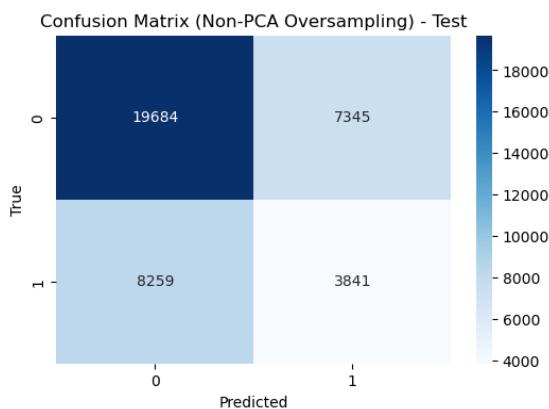
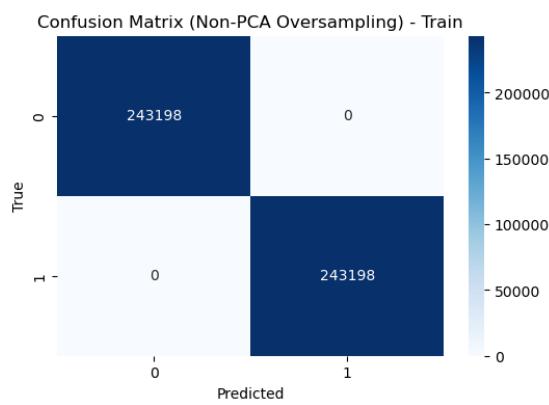
Berikut adalah hasil prediksi untuk *training* dan *testing*:



Hasil *Random Forest* yang dilakukan dengan *oversampling* tanpa PCA dengan parameter terbaik yaitu:

```
{'n_estimators': 200,
 'min_samples_split': 2,
 'min_samples_leaf': 1,
 'max_depth': None,
 }.
```

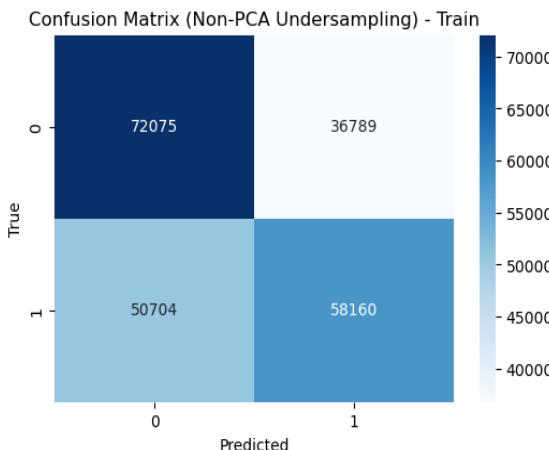
Berikut adalah hasil prediksi untuk *training* dan *testing*:

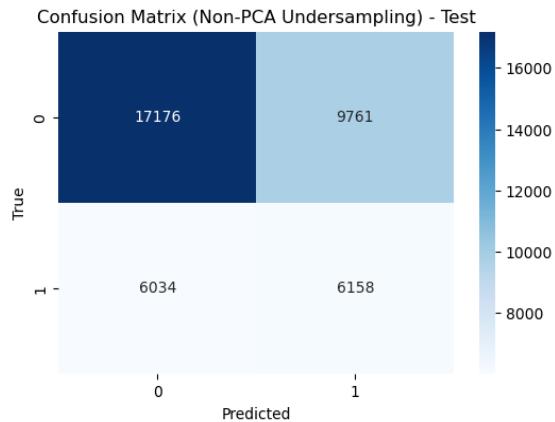


Hasil *Random Forest* yang dilakukan dengan *undersampling* tanpa PCA dengan parameter terbaik yaitu:

```
{'n_estimators': 100,
 'min_samples_split': 2,
 'min_samples_leaf': 2,
 'max_depth': 10,
 }.
```

Berikut adalah hasil prediksi untuk *training* dan *testing*:

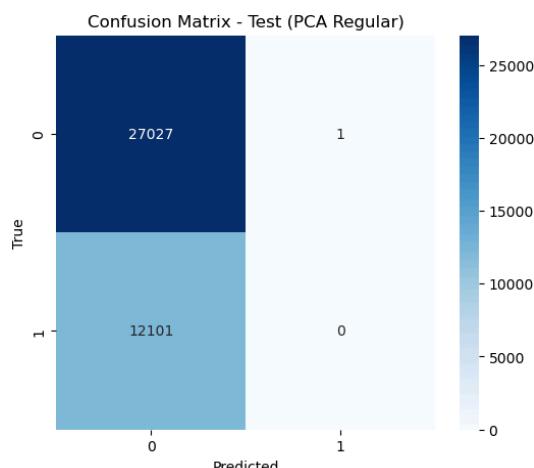
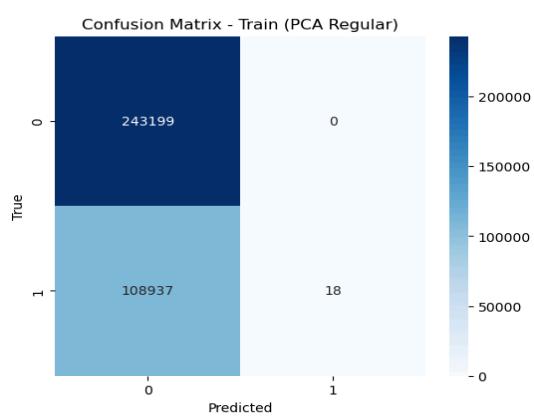




Hasil *Random Forest regular* yang diproses menggunakan PCA dengan parameter terbaik yaitu:

```
{'n_estimators': 100,
'min_samples_split': 2,
'min_samples_leaf': 2,
'max_depth': 10,
}.
```

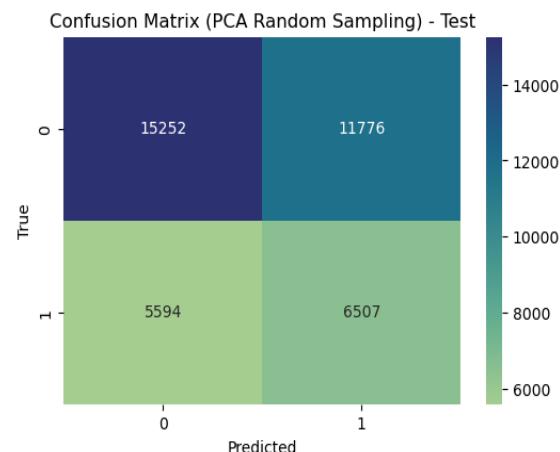
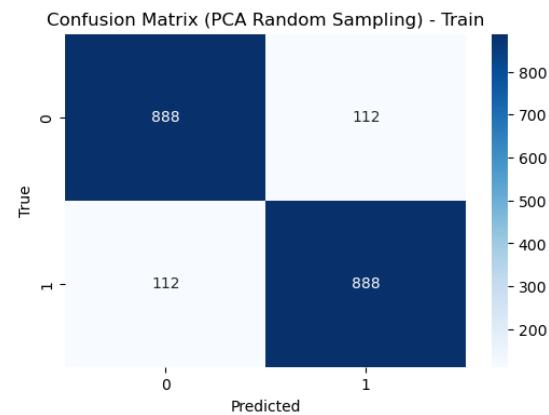
Berikut adalah hasil prediksi untuk *training* dan *testing*:



Hasil *Random Forest* yang diproses menggunakan PCA dan *random sampling* dengan parameter terbaik yaitu:

```
{'n_estimators': 100,
'min_samples_split': 5,
'min_samples_leaf': 1,
'max_depth': 10,
}.
```

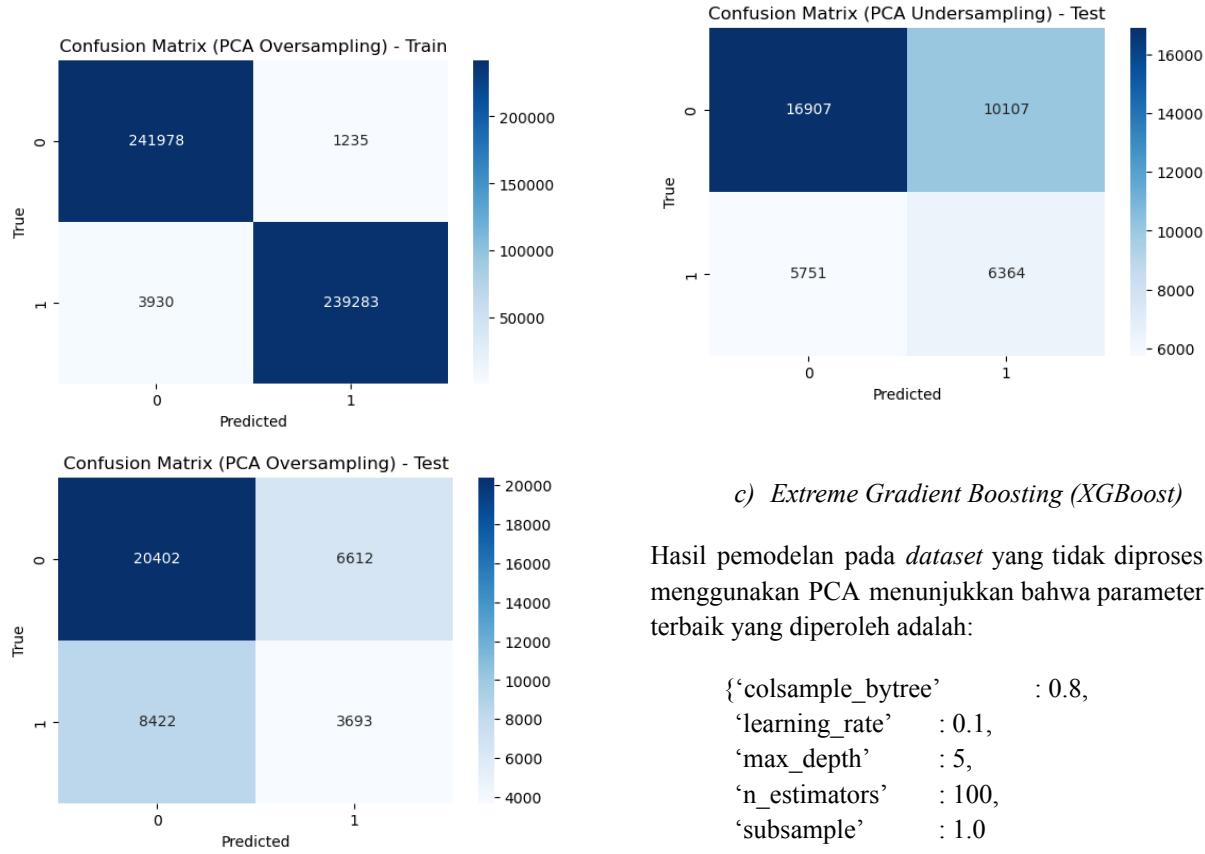
Berikut adalah hasil prediksi untuk *training* dan *testing*:



Hasil *Random Forest* yang diproses menggunakan PCA dan *oversampling* dengan parameter terbaik yaitu:

```
{'n_estimators': 200,
'min_samples_split': 5,
'min_samples_leaf': 1,
'max_depth': None,
}.
```

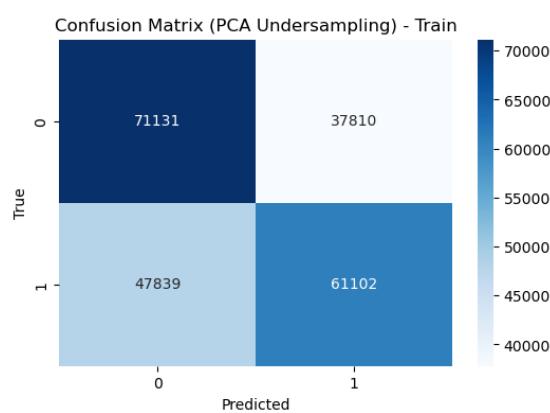
Berikut adalah hasil prediksi untuk *training* dan *testing*:



Hasil *Random Forest* yang diproses menggunakan PCA dan *undersampling* dengan parameter terbaik yaitu:

```
{'n_estimators': 200,
 'min_samples_split': 5,
 'min_samples_leaf': 1,
 'max_depth': 10,
}.
```

Berikut adalah hasil prediksi untuk *training* dan *testing*:

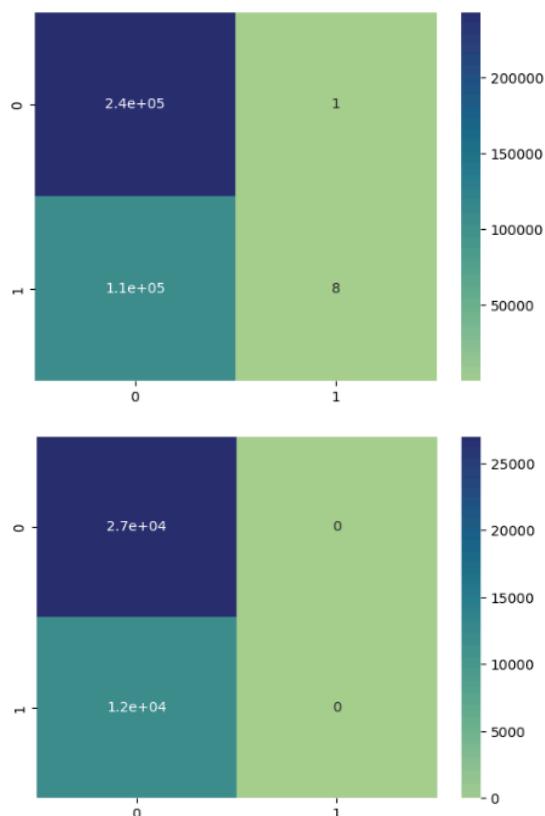


### c) Extreme Gradient Boosting (XGBoost)

Hasil pemodelan pada *dataset* yang tidak diproses menggunakan PCA menunjukkan bahwa parameter terbaik yang diperoleh adalah:

```
{'colsample_bytree': 0.8,
 'learning_rate': 0.1,
 'max_depth': 5,
 'n_estimators': 100,
 'subsample': 1.0
}.
```

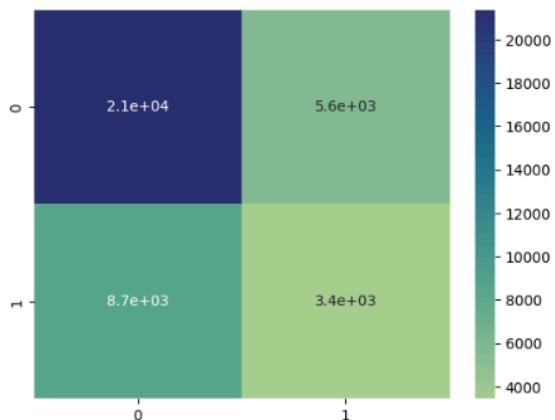
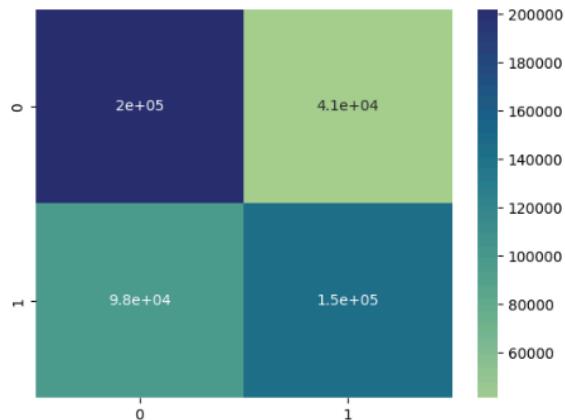
Selain itu, berikut adalah hasil prediksi untuk data *training* dan *testing*:



Hasil pemodelan pada *dataset* yang tidak diproses menggunakan PCA, tetapi telah dilakukan *over-sampling*, menunjukkan bahwa parameter terbaik yang diperoleh adalah:

```
{'colsample_bytree' : 0.7,
 'learning_rate' : 0.2,
 'max_depth' : 7,
 'n_estimators' : 500,
 'subsample' : 1.0
}.
```

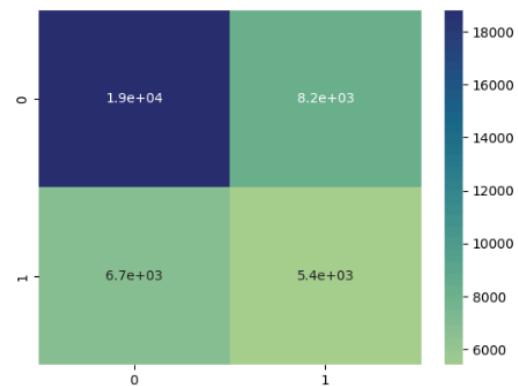
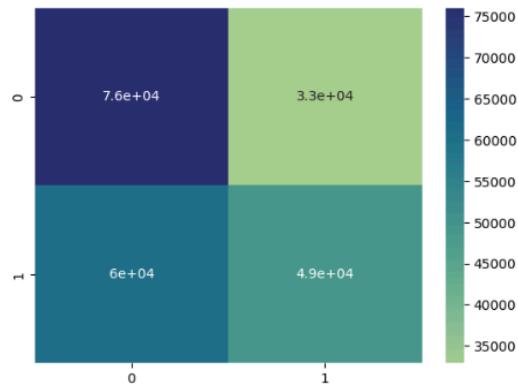
Selain itu, berikut adalah hasil prediksi untuk data training dan testing:



Hasil pemodelan pada *dataset* yang tidak diproses menggunakan PCA, tetapi telah dilakukan *under-sampling*, menunjukkan bahwa parameter terbaik yang diperoleh adalah:

```
{'colsample_bytree' : 1.0,
 'learning_rate' : 0.01,
 'max_depth' : 3,
 'n_estimators' : 200,
 'subsample' : 0.7
}.
```

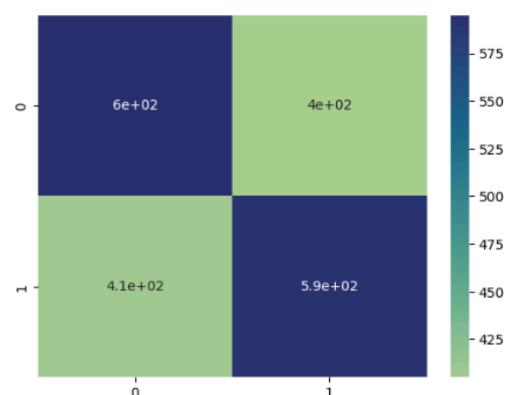
Selain itu, berikut adalah hasil prediksi untuk data training dan testing:

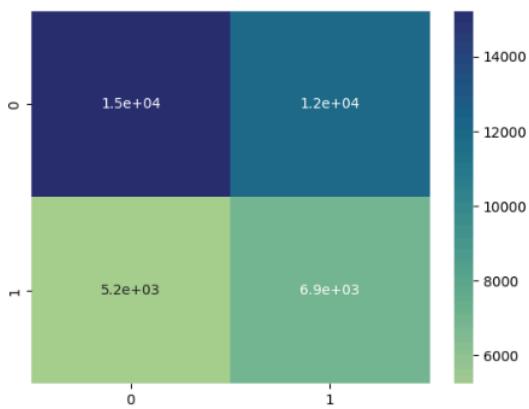


Hasil pemodelan pada *dataset* yang tidak diproses menggunakan PCA, tetapi telah dilakukan *random sampling* sebanyak 1000 dari setiap kelas, menunjukkan bahwa parameter terbaik yang diperoleh adalah:

```
{'colsample_bytree' : 0.8,
 'learning_rate' : 0.01,
 'max_depth' : 3,
 'n_estimators' : 100,
 'subsample' : 0.7
}.
```

Selain itu, berikut adalah hasil prediksi untuk data training dan testing:

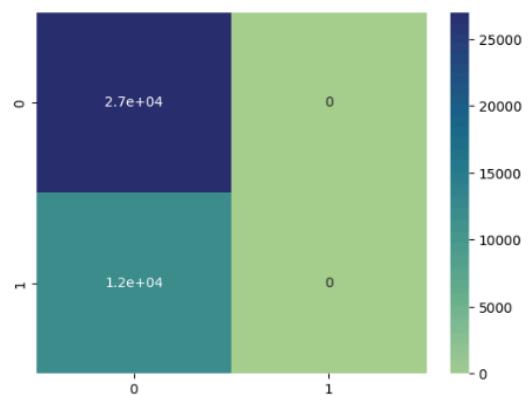
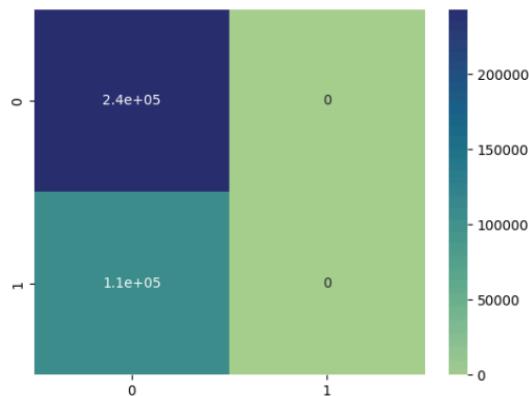




Hasil pemodelan pada *dataset* yang diproses menggunakan PCA menunjukkan bahwa parameter terbaik yang diperoleh adalah:

```
{'colsample_bytree' : 0.7,
 'learning_rate' : 0.01,
 'max_depth' : 3,
 'n_estimators' : 100,
 'subsample' : 0.7
}.
```

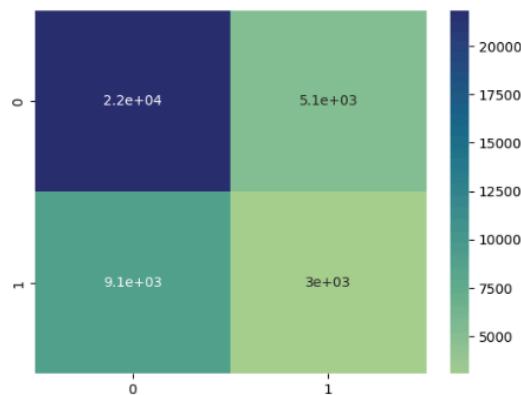
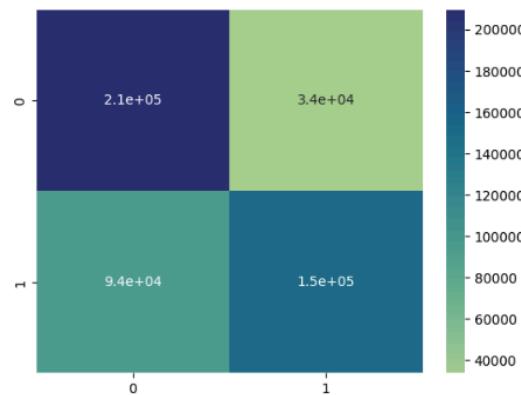
Selain itu, berikut adalah hasil prediksi untuk data *training* dan *testing*:



Hasil pemodelan pada *dataset* yang diproses menggunakan PCA dan *oversampling* menunjukkan bahwa parameter terbaik yang diperoleh adalah:

```
{'colsample_bytree' : 1.0,
 'learning_rate' : 0.2,
 'max_depth' : 7,
 'n_estimators' : 500,
 'subsample' : 1.0
}.
```

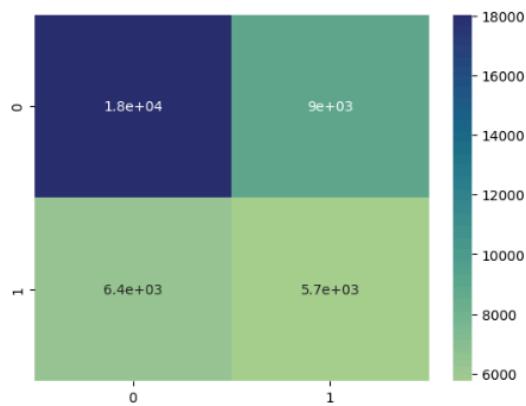
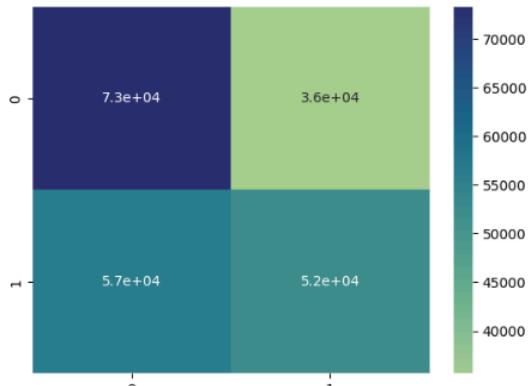
Selain itu, berikut adalah hasil prediksi untuk data *training* dan *testing*:



Hasil pemodelan pada *dataset* yang diproses menggunakan PCA dan *under-sampling* menunjukkan bahwa parameter terbaik yang diperoleh adalah:

```
{'colsample_bytree' : 0.8,
 'learning_rate' : 0.01,
 'max_depth' : 5,
 'n_estimators' : 200,
 'subsample' : 0.8
}.
```

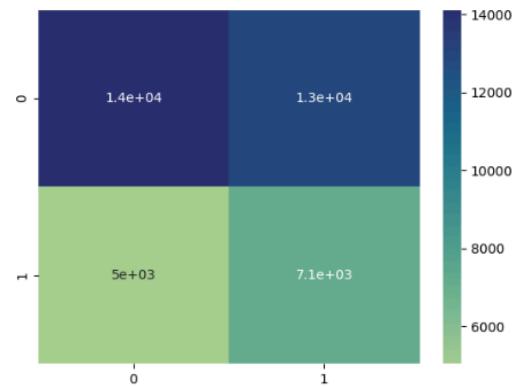
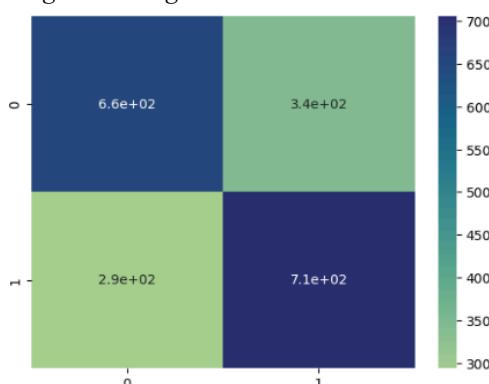
Selain itu, berikut adalah hasil prediksi untuk data *training* dan *testing*:



Hasil pemodelan pada *dataset* yang diproses menggunakan PCA dan *random sampling* sebanyak 1000 dari setiap kelas menunjukkan bahwa parameter terbaik yang diperoleh adalah:

```
{'colsample_bytree' : 0.8,
 'learning_rate' : 0.01,
 'max_depth' : 5,
 'n_estimators' : 100,
 'subsample' : 1.0
}.
```

Selain itu, berikut adalah hasil prediksi untuk data *training* dan *testing*:



## B. Evaluasi

### 1) Support Vector Machine (SVM)

Hasil **akurasi data *training*** dengan SVM adalah sebagai berikut:

Tipe Model	Balanced Accuracy	F1-Score
SVM Reguler Non-PCA	0.5	0
SVM Random Sampling Non-PCA	0.575	0.5360
SVM Undersampling Non-PCA	0.5728	0.5279
SVM Reguler PCA	0.5	0
SVM Random Sampling PCA	0.558	0.5196
SVM Undersampling PCA	0.5723	0.5281

Hasil **akurasi data *testing*** dengan SVM adalah sebagai berikut:

Tipe Model	Balanced Accuracy	F1-Score
SVM Reguler Non-PCA	0.5	0

SVM Random Sampling Non-PCA	0.5697	0.4292
SVM Undersampling Non-PCA	0.5739	0.4332
SVM Reguler PCA	0.5	0
SVM Random Sampling PCA	0.5709	0.4267
SVM Undersampling PCA	0.5702	0.4274

Berdasarkan hasil evaluasi, model terbaik secara keseluruhan adalah **SVM Undersampled Non-PCA**, yang unggul dalam beberapa metrik utama. Pada data pelatihan, *Balanced Accuracy* tercatat sebesar 0.57 dan *F1-Score* sebesar 0.53, sementara pada data pengujian, *Balanced Accuracy* mencapai 0.57 dan *F1-Score* sebesar 0.43. Meskipun nilai *F1-Score* pada data pengujian tidak setinggi yang diinginkan, performa model ini tetap menunjukkan kemampuan yang lebih baik dalam menangani ketidakseimbangan kelas, terutama dibandingkan dengan model lainnya.

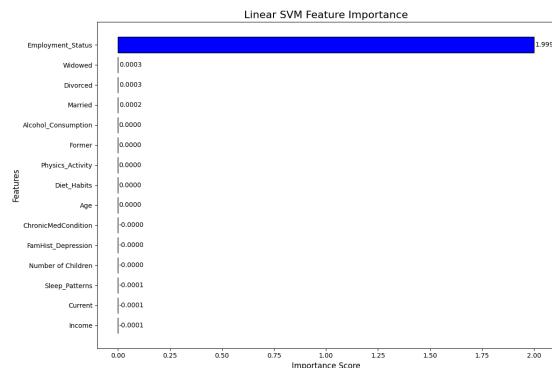
Pendekatan SVM *Undersampled* Non-PCA juga memberikan keuntungan dari segi kesederhanaan dan efisiensi. Model ini menggunakan teknik undersampling untuk menangani ketidakseimbangan kelas, tanpa memerlukan transformasi PCA atau teknik *resampling* lain. Hal ini memungkinkan model untuk tetap efisien dalam hal waktu dan sumber daya, sambil tetap mempertahankan akurasi yang cukup baik pada data pelatihan dan pengujian.

Sementara itu, teknik transformasi PCA tidak memberikan peningkatan signifikan pada performa model SVM untuk dataset ini. Baik pada data pelatihan maupun pengujian, hasil yang diperoleh dari model *Undersampled* Non-PCA lebih konsisten dan lebih baik dibandingkan dengan model yang menggunakan PCA, baik secara *regular* maupun *undersampled*.

Kesimpulannya, SVM *Undersampled* Non-PCA adalah model yang lebih andal dalam mendeteksi

depresi untuk penelitian ini. Dengan performa yang lebih seimbang dan kemampuan untuk menangani ketidakseimbangan kelas, pendekatan ini menjadi pilihan terbaik untuk digunakan pada dataset serupa di masa depan.

Selanjutnya adalah grafik *feature importance* yang menunjukkan fitur mana yang paling berdampak:



Berdasarkan grafik *feature importance* ini, fitur-fitur yang mempengaruhi terhadap keberadaan depresi adalah status pekerjaan.

Status pekerjaan punya pengaruh besar terhadap risiko depresi. Orang yang memiliki pekerjaan umumnya punya stabilitas finansial dan struktur rutinitas yang membantu menjaga kesehatan mental. Tapi jika orang kehilangan pekerjaan atau ada ketidakpastian kerja, depresi mulai muncul. Salah satu studi dari Pubmed menunjukkan bahwa pengangguran atau perubahan status pekerjaan secara signifikan dapat meningkatkan risiko gangguan mental termasuk depresi. Faktor-faktor seperti kehilangan pekerjaan, tekanan finansial, dan ketidakpastian terkait pekerjaan sering dikaitkan dengan kondisi mental yang memburuk, terutama depresi. [26] [27]

## 2) Random Forest (RF)

Hasil **akurasi data training** dengan *Random Forest* adalah sebagai berikut:

Tipe Model	*Balanced Accuracy	*F1-Score
Random Forest Reguler Non-PCA	0.5000	0.5642
Random Forest Random Sampling	0.9005	0.9005

Non-PCA		
Random Forest Oversampling Non-PCA	1.0000	1.0000
Random Forest Undersampling Non-PCA	0.5982	0.5965
Random Forest Reguler PCA	0.5001	0.5643
Random Forest Random Sampling PCA	0.8880	0.8880
Random Forest Oversampling PCA	0.9894	0.9894
Random Forest Undersampling PCA	0.6069	0.6061

\*Catatan: angka dibulatkan ke 4 angka belakang desimal.

Hasil **akurasi data testing** dengan *Random Forest* adalah sebagai berikut:

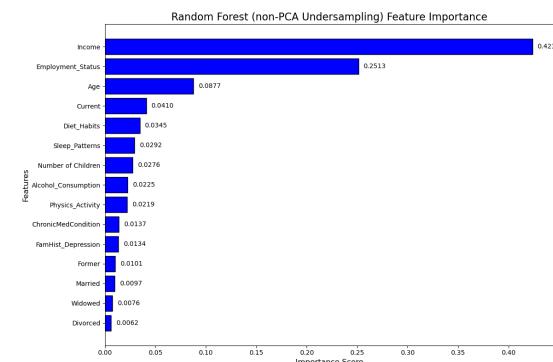
Tipe Model	*Balanced Accuracy	*F1-Score
Random Forest Reguler Non-PCA	0.5000	0.5644
Random Forest Random Sampling Non-PCA	0.5554	0.5636
Random Forest Oversampling Non-PCA	0.528	0.5967
Random Forest Undersampling Non-PCA	0.5714	0.6081
Random Forest Reguler PCA	0.5000	0.5644
Random Forest Random Sampling PCA	0.5510	0.5726
Random Forest Oversampling PCA	0.5300	0.6065

Random Forest Undersampling PCA	0.5756	0.6078
------------------------------------	--------	--------

\*Catatan: angka dibulatkan ke 4 angka belakang desimal.

Berdasarkan hasil pelatihan model, model terbaik secara keseluruhan adalah **Random Forest Non-PCA Undersampling**. Model ini hampir setara dengan PCA *Undersampling*, tetapi F1-score pada algoritma ini lebih berbobot daripada *Balanced Accuracy*. Skor F1 adalah metrik yang lebih komprehensif sebagai ukuran keseimbangan prediksi antara kelas positif dan negatif, yang penting untuk skenario seperti deteksi depresi guna meminimalisir positif palsu (*False Positives*) dan negatif palsu (*False Negatives*). Sementara, *Balanced Accuracy* efektif dalam menangani ketidakseimbangan kelas, skor F1 yang lebih tinggi dari model Non-PCA *Undersampling* mengindikasikan model ini mencapai keseimbangan keseluruhan yang lebih baik dalam prediksi kelas. Kemampuan model untuk menggeneralisasi dengan baik di seluruh data *training* dan *testing* semakin mendukung kinerjanya yang kuat, menjadikannya pilihan terbaik di antara metode yang dievaluasi. Meskipun demikian, skor *Balanced Accuracy* yang lebih rendah masih dalam jangkauan yang dapat diterima dengan perbedaan 0.0042, sehingga masih serupa dengan PCA *Undersampling*.

Selanjutnya adalah grafik *feature importance* yang menunjukkan fitur mana yang paling berdampak:



Berdasarkan grafik *feature importance* ini, fitur-fitur yang berpengaruh terhadap keberadaan depresi adalah jumlah gaji, status pekerjaan, dan usia. Hal ini termasuk dalam kategori *stres finansial* (Financial Stress). Faktor penyebab ini cukup umum untuk departemen pemerintah dari

Australia[23] atau Inggris[24] untuk mengakui dan memperhatikan masalah ini. Memiliki pekerjaan yang stabil dengan penghasilan yang stabil, terutama di usia dini, merupakan salah satu solusi untuk mengurangi depresi dengan jumlah yang signifikan.

### 3) Extreme Gradient Boosting (XGBoost)

Hasil **akurasi data training** pemodelan dengan algoritma XGBoost, sebagai berikut:

Tipe Model	*Balanced Accuracy	*F1-Score
XGBoost Reguler Non-PCA	0.5000	0.5644
XGBoost Oversampled Non-PCA	0.7132	0.7092
XGBoost Undersampled Non-PCA	0.5731	0.5663
XGBoost Random Sampling Non-PCA	0.594	0.594
XGBoost Reguler PCA	0.5	0.5643
XGBoost Oversampled PCA	0.7273	0.7332
XGBoost Undersampled PCA	0.5769	0.5729
XGBoost Random Sampling PCA	0.681	0.6808

\*Catatan: angka dibulatkan ke 4 angka belakang desimal.

Hasil **akurasi data testing** pemodelan dengan algoritma XGBoost, sebagai berikut:

Tipe Model	*Balanced Accuracy	*F1-Score
XGBoost Reguler Non-PCA	0.5	0.8165
XGBoost Oversampled	0.5376	0.6506

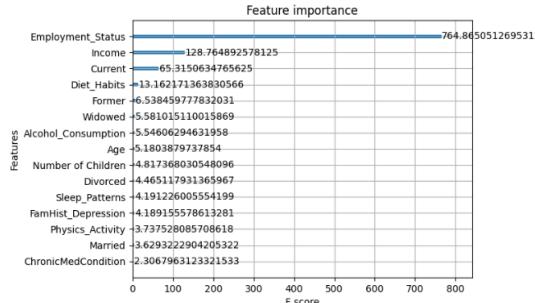
Non-PCA		
XGBoost <i>Undersampled</i> Non-PCA	0.5712	0.6135
XGBoost <i>Random Sampling</i> Non-PCA	0.5660	0.5489
XGBoost <i>Reguler</i> PCA	0.5	0.8165
XGBoost <i>Oversampled</i> PCA	0.5304	0.6135
XGBoost <i>Undersampled</i> PCA	0.5708	0.5988
XGBoost <i>Random Sampling</i> PCA	0.5540	0.5252

\*Catatan: angka dibulatkan ke 4 angka belakang desimal.

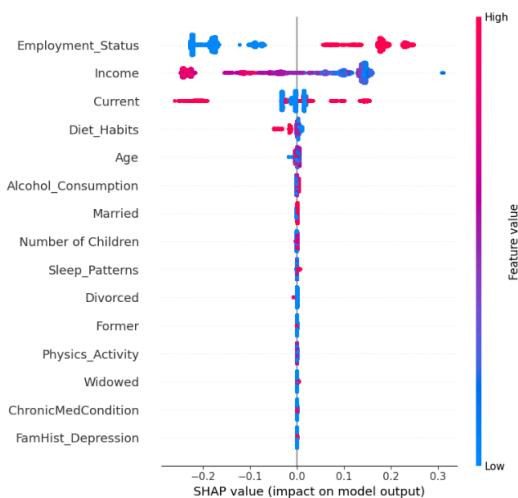
Pada evaluasi performa model XGBoost, hasil terbaik diperoleh oleh model **XGBoost Undersampled Non-PCA**, karena model ini menunjukkan kinerja yang lebih seimbang dalam memprediksi kedua kelas target, dengan **Balanced Accuracy tertinggi** pada data testing sebesar **0.5712**, yang mengindikasikan kemampuan yang lebih baik dalam menangani ketidakseimbangan kelas pada data. Walaupun *F1-Score* model ini tidak setinggi beberapa model lainnya, konsistensi performa *Balanced Accuracy* antara data *training* dan data *testing* menunjukkan bahwa model ini cenderung lebih stabil dalam memprediksi keberadaan depresi tanpa adanya *overfitting*, dalam arti mempelajari pola-pola data beserta *noise* dalam data *training*. Namun, perlu diperhatikan bahwa model ini kurang baik dalam memprediksi depresi, terutama karena memperoleh *balanced accuracy* yang cenderung rendah, yaitu dibawah 80%. Hal ini menunjukkan bahwa walaupun algoritma/model secara teoritis dapat memprediksi kedua kelas lebih seimbang dibandingkan algoritma lainnya, tingkat akurasi dalam mendeteksi kasus depresi masih terbatas sehingga hasil prediksi perlu ditingkatkan kembali untuk mencapai performa yang lebih andal.

Sebagai tambahan, model ini memiliki metrik kekuatan signifikansi fitur yang membantu dalam memahami faktor-faktor yang mempengaruhi

prediksi keberadaan depresi. Berikut adalah rincian *feature importance* dan SHAP (*SHapley Additive exPlanations*) *values* yang dapat mendukung interpretasi lebih lanjut terhadap pengaruh dalam prediksi tersebut.



Berdasarkan grafik *feature importance* ini, fitur yang berpengaruh terhadap prediksi model adalah status kerja, pemasukan, status perokok aktif, kualitas pola makan, status mantan perokok.



Berdasarkan kedua grafik di atas, fitur status kerja, pemasukan/gaji, status aktif merokok, dan pola makan merupakan empat faktor utama bagi model **XGBoost Undersampled Non-PCA** dalam menentukan status depresi seseorang, karena memiliki *feature importance* (ukuran seberapa signifikansi suatu fitur terhadap hasil prediksi pemodelan) yang lebih tinggi dibandingkan fitur-fitur lainnya. Akan tetapi, berdasarkan hasil visualisasi SHAP *values*, yang pada umumnya digunakan untuk memahami cara algoritma membuat prediksi, ditemukan bahwa orang yang tidak memiliki pekerjaan (*Employment\_Status* = 0) cenderung memiliki risiko depresi yang lebih rendah, sedangkan mereka yang memiliki

pekerjaan (*Employment\_Status* = 1) cenderung lebih berisiko terkena depresi.

Namun, terdapat kontradiksi antara fitur ini dan fitur pemasukan/gaji seseorang, di mana gaji rendah cenderung meningkatkan risiko depresi, sementara pemasukan tinggi menurunkan risiko depresi. Untuk fitur status aktif merokok, hasil menunjukkan bahwa orang yang merokok memiliki pengaruh terhadap status depresi, tetapi model tidak dapat memprediksi dengan jelas apakah resikonya meningkat atau menurun, karena plot tersebut pada sumbu positif dan negatif. Terakhir, dari fitur pola makan, terlihat bahwa seseorang dengan pola makan yang sehat (*Diet\_Habits* = 3, *Healthy*) sedikit menurunkan risiko terkena depresi.

### C. Kesimpulan

Berdasarkan hasil evaluasi, **XGBoost** merupakan model terbaik dalam skenario **Undersampled tanpa PCA**, walaupun memiliki akurasi lebih tinggi sedikit daripada **Random Forest Undersampled tanpa PCA**. Model ini memiliki *balance accuracy* dan *F1-score* terbaik dengan masing-masing sebesar **0.5712** dan **0.6135**, menunjukkan kemampuan yang lebih baik dalam menyeimbangkan presisi dan *recall* untuk data yang tidak seimbang. Meskipun *balanced accuracy*-nya (0.5712) hampir sama dengan model lainnya, *F1-score* yang lebih tinggi menjadikannya pilihan terbaik untuk memprediksi *mental illness* secara lebih andal. Empat fitur utama yang paling berkontribusi terhadap prediksi ini adalah status kerja (orang yang memiliki pekerjaan cenderung memiliki risiko depresi yang tinggi), pemasukan (gaji yang rendah meningkatkan risiko depresi), status aktif merokok, pola makan (pola makan sehat cenderung menurunkan risiko depresi). Namun, karena algoritma prediksi secara keseluruhan belum menunjukkan performa yang baik, perlu dicatat bahwa penilaian risiko depresi seharusnya tidak hanya didasarkan pada faktor demografi, pola hidup, dan ekonomi, tetapi juga mempertimbangkan faktor biologis yang dapat memberikan gambaran lebih menyeluruh. Misalnya, kadar hormon tertentu seperti serotonin dan kortisol, serta kondisi medis lainnya yang dapat berpotensi mempengaruhi risiko seseorang dalam mengalami depresi. Dengan mempertimbangkan faktor-faktor biologis, studi dapat memperkaya analisis dan memberikan hasil pemodelan

algoritma pembelajaran mesin yang lebih akurat dan andal dalam memahami risiko depresi.

## Daftar Pustaka

- [1] “Depressive disorder (depression).” World Health Organization, Mar. 31, 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/depression>
- [2] World Health Organization, *World health statistics 2019: monitoring health for the SDGs, sustainable development goals*. Geneva: World Health Organization, 2019. Accessed: Sep. 19, 2024. [Online]. Available: <https://iris.who.int/handle/10665/324835>
- [3] O. Remes, J. F. Mendes, and P. Templeton, “Biological, Psychological, and Social Determinants of Depression: A Review of Recent Literature,” *Brain Sci.*, vol. 11, no. 12, p. 1633, Dec. 2021, doi: 10.3390/brainsci11121633.
- [4] R. Stanton *et al.*, “Depression, Anxiety and Stress during COVID-19: Associations with Changes in Physical Activity, Sleep, Tobacco and Alcohol Use in Australian Adults,” *Int. J. Environ. Res. Public. Health*, vol. 17, no. 11, p. 4065, Jun. 2020, doi: 10.3390/ijerph17114065.
- [5] N. Craddock and L. Mynors-Wallis, “Psychiatric diagnosis: impersonal, imperfect and important,” *Br. J. Psychiatry*, vol. 204, no. 2, pp. 93–95, Feb. 2014, doi: 10.1192/bjp.bp.113.133090.
- [6] Ö. ÇeliK, “A Research on Machine Learning Methods and Its Applications,” *J. Educ. Technol. Online Learn.*, vol. 1, no. 3, pp. 25–40, Sep. 2018, doi: 10.31681/jetol.457046.
- [7] J. Jia, “70 Accuracy Random Forest.” [Online]. Available: <https://www.kaggle.com/code/jasoncoderjia/70-accuracy-random-forest>
- [8] Sadik, “Depression Prediction With Feature Selection.” [Online]. Available: <https://www.kaggle.com/code/renjiabarai/depression-prediction-with-feature-selection#Data-Preparation>
- [9] S. Aleem, N. U. Huda, R. Amin, S. Khalid, S. S. Alshamrani, and A. Alshehri, “Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions,” *Electronics*, vol. 11, no. 7, p. 1111, Mar. 2022, doi: 10.3390/electronics11071111.
- [10] “Depression Dataset.” Accessed: Sep. 18, 2024. [Online]. Available: <https://www.kaggle.com/datasets/anthonytherrien/depression-dataset>
- [11] M. D. Sacchet, G. Prasad, L. C. Foland-Ross, P. M. Thompson, and I. H. Gotlib, “Support Vector Machine Classification of Major Depressive Disorder Using Diffusion-Weighted Neuroimaging and Graph Theory,” *Front. Psychiatry*, vol. 6, Feb. 2015, doi: 10.3389/fpsyg.2015.00021.
- [12] L. Su, L. Wang, H. Shen, G. Feng, and D. Hu, “Discriminative analysis of non-linear brain connectivity in schizophrenia: an fMRI Study,” *Front. Hum. Neurosci.*, vol. 7, Oct. 2013, doi: 10.3389/fnhum.2013.00702.
- [13] “Random Forest Algorithm - How It Works and Why It Is So Effective.” Accessed: Sep. 21, 2024. [Online]. Available: <https://www.turing.com/kb/random-forest-algorithm>
- [14] M. Schott, “Random Forest Algorithm for Machine Learning,” Capital One Tech. Accessed: Sep. 30, 2024. [Online]. Available: <https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8c9feb>
- [15] Y. Xin and X. Ren, “Predicting depression among rural and urban disabled elderly in China using a random forest classifier,” *BMC Psychiatry*, vol. 22, no. 1, p. 118, Feb. 2022, doi: 10.1186/s12888-022-03742-4.
- [16] M. Srinivas, G. Sucharitha, and A. Matta, *Machine Learning Algorithms and Applications*. Newark: John Wiley & Sons, Incorporated, 2021.
- [17] A. Srinivasaraghavan and V. Joseph, *Machine Learning*, 1st ed. Wiley India Pvt. Ltd., 2019.
- [18] A. Sharma and W. J. M. I. Verbeke, “Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081),” *Front. Big Data*, vol. 3, p. 15, Apr. 2020, doi: 10.3389/fdata.2020.00015.
- [19] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, 6th ed. Upper Saddle River, N.J: Pearson Prentice Hall, 2007.
- [20] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 5th ed. Chapman and Hall/CRC, 2020. doi: 10.1201/9780429186196.
- [21] M. Sriningsih, D. Hatidja, and J. D. Prang, “PENANGANAN MULTIKOLINEARITAS DENGAN MENGGUNAKAN ANALISIS REGRESI KOMPONEN UTAMA PADA KASUS IMPOR BERAS DI PROVINSI SULUT,” *J. Ilm. SAINS*, vol. 18, no. 1, p. 18, Jul. 2018, doi: 10.35799/jis.18.1.2018.19396.
- [22] H. Abdi and L. J. Williams, “Principal component analysis,” *WIREs Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.
- [23] C. C. scheme=AGLSTERMS. AglsAgent; corporateName=Department of Education;

- address=50 Marcus Clarke St, "Financial stress." Accessed: Nov. 17, 2024. [Online]. Available:  
<https://www.education.gov.au/integrated-data-research/benefits-educational-attainment/financial-stress>
- [24] "Coping with financial worries," nhs.uk. Accessed: Nov. 17, 2024. [Online]. Available:  
<https://www.nhs.uk/mental-health/advice-for-life-situations-and-events/how-to-cope-with-financial-worries/>
- [25] H. Xie *et al.*, "Parkinson's disease with mild cognitive impairment may have a lower risk of cognitive decline after subthalamic nucleus deep brain stimulation: a retrospective cohort study," *Frontiers in Human Neuroscience*, vol. 16, 2022. [Online]. Available:  
<https://www.frontiersin.org/articles/10.3389/fnhum.2022.943472/full>.
- [26] C. K. W. Lai, W. C. Beasley, T. Kim, K. Marsh, S. E. Hofmann, and D. S. Yeung, "Evidence-Based Approaches to Dementia Prevention," *Nature Reviews Neurology*, vol. 7, no. 10, pp. 482–493, Oct. 2011. [Online]. Available:  
<https://pubmed.ncbi.nlm.nih.gov/21806873/>
- [27] T. Liu, J. Qiu, X. Zhao, S. Fan, Y. Wang, and Z. Liu, "Advancements in Machine Learning for Alzheimer's Disease Prediction," *Journal of Alzheimer's Disease*, vol. 60, no. 2, pp. 321–333, Nov. 2017. [Online]. Available:  
<https://pubmed.ncbi.nlm.nih.gov/29207885/>