Laporan Akhir Tugas MK. Data Mining (KOM1338), Semester Genap 2024/2025
Program Studi Sarjana Ilmu Komputer
SSMI IPB

Perbandingan Model Logistic Regression, Support Vector Machine, dan XGBoost dalam Klasifikasi Depresi Siswa

Anisa Hayatullah (G6401221009), Roshan Zakaria (G6401221010), Sri Arini Ismayasari (G6401221029), Nurul Fadillah (G6401221078) [1*]
Kelompok: 3, Kelas Paralel: 2

Abstrak

Kesehatan mental merupakan isu yang serius secara global. Kegagalan dalam menangani masalah kesehatan mental berpotensi memberikan dampak jangka panjang pada kesehatan fisik dan mental kedepannya. Hal tersebut membuat deteksi dini terhadap risiko gangguan mental menjadi sangat penting. *Machine learning* dapat digunakan untuk mendeteksi risiko adanya indikasi gangguan mental pada individu. Penelitian ini menggunakan *machine learning* dengan beberapa metode seperti Logistic Regression, SVM, dan XGBoost untuk mengklasifikasikan depresi pada siswa. Data yang digunakan merupakan data sekunder yang berasal dari platform Kaggle. Hasil menunjukkan bahwa metode terbaik dalam kasus ini adalah SVM dengan kernel sigmoid yang memiliki akurasi prediksi sebesar 84,49946% dengan nilai *precision, recall, dan f1-score* yang cukup tinggi, serta perolehan *learning curve* yang lebih stabil dibandingkan metode Logistic Regression dan XGBoost.

Kata Kunci**:** depresi, *machine learning*, kesehatan mental, klasifikasi.

PENDAHULUAN

Latar Belakang

Masalah kesehatan mental menjadi isu serius yang berdampak besar terhadap kondisi emosional, cara berpikir, serta kemampuan seseorang dalam berinteraksi sosial (Chung dan Teo 2022). Data terbaru Organisasi Kesehatan Dunia (WHO 2024) mengungkapkan bahwa secara global, 1 dari 7 remaja berusia 10 hingga 19 tahun menderita gangguan mental, namun kasus ini masih belum terdeteksi secara tepat. Masalah psikologis seperti depresi, gangguan kecemasan, dan perilaku *disruptive* tercatat sebagai kontributor utama beban penyakit dan disabilitas pada populasi remaja. Hal yang lebih mengkhawatirkan lagi, tindakan bunuh diri menempati peringkat ketiga penyebab kematian pada kelompok usia produktif 15-29 tahun.

Kegagalan dalam menangani masalah kesehatan mental pada usia muda berpotensi memberikan dampak jangka panjang hingga masa dewasa yang tidak hanya memengaruhi kesehatan fisik dan mental, tetapi juga membatasi peluang individu untuk menjalani kehidupan yang produktif dan bermakna (Garriga R *et al.* 2022). Dalam praktiknya, permintaan terhadap pelayanan kesehatan jiwa umumnya meningkat saat pasien mengalami krisis, yaitu ketika mereka tidak mampu mengurus diri sendiri, berfungsi dalam

[1]Program Studi Sarjana Ilmu Komputer, Sekolah Sains Data, Matematika dan Informatika (SSMI), Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Sarjana Ilmu Komputer, SSMI IPB; Surel: annanisa@apps.ipb.ac.id, roshanzakaria@apps.ipb.ac.id, sriarini@apps.ipb.ac.id, nrifdiihnurul@apps.ipb.ac.id

masyarakat atau bahkan menunjukkan potensi untuk menyakiti diri sendiri maupun orang lain (Paton *et al.* 2016). Kebiasaan ini menghambat upaya pencegahan dilakukan secara optimal karena pasien telah memasuki kondisi krisis (Horwitz *et al.* 2019).

Oleh karena itu, deteksi dini terhadap risiko gangguan mental menjadi sangat krusial dalam upaya meningkatkan hasil perawatan dan efisiensi pengelolaan sumber daya pelayanan kesehatan jiwa (Chung dan Teo 2022). Glaz *et al.* (2021) menyebutkan dalam konteks ini, penerapan teknologi seperti *machine learning* menjadi solusi potensial yang dapat membantu dalam mengidentifikasi individu dengan risiko gangguan mental secara lebih cepat dan tepat. Kemampuan *machine learning* dalam mengolah data untuk membentuk sistem cerdas melalui pola dari data historis, serta memprediksi atau mengklasifikasikan kondisi baru dapat dimanfaatkan untuk membantu proses diagnosis dini terhadap permasalahan kesehatan mental (Iyortsuun *et al.* 2022).

Sejumlah studi menunjukkan bahwa algoritma m*achine learning* telah digunakan secara luas dalam bidang kesehatan mental, khususnya untuk membantu proses klasifikasi dan prediksi kondisi mental seseorang. Logistic Regression, Support Vector Machine (SVM), dan XGBoost merupakan tiga algoritma yang sering diterapkan dalam tugas klasifikasi, masing-masing menawarkan pendekatan yang berbeda dalam mengolah data dan membentuk model prediktif. Melalui perbandingan kinerja ketiga model ini dalam mengklasifikasikan risiko depresi pada siswa diharapkan dapat memberikan wawasan mengenai metode yang paling tepat dan efektif dalam mengidentifikasi risiko depresi pada siswa, sehingga dapat mendukung proses deteksi dini dan intervensi secara lebih responsif.

Tujuan

Tugas akhir ini bertujuan untuk menganalisis dan membandingkan kinerja tiga algoritma machine learning, yaitu Logistic Regression, Support Vector Machine (SVM), dan XGBoost, dalam membangun model klasifikasi untuk mendeteksi risiko depresi pada siswa. Perbandingan dilakukan untuk menentukan algoritma yang memberikan hasil paling optimal dalam hal akurasi, efisiensi, dan kemampuan generalisasi. Hasil dari analisis ini diharapkan dapat menjadi dasar dalam pemilihan metode yang tepat untuk deteksi dini gangguan kesehatan mental siswa. Dengan pendekatan yang lebih akurat dan efisien, deteksi dini dapat membantu upaya pencegahan dan penanganan yang lebih optimal.

Ruang Lingkup

Penelitian ini berfokus pada analisis klasifikasi status depresi siswa dan evaluasi performa model berdasarkan data yang tersedia. Tahapan yang dilakukan meliputi praproses data, pembagian data menjadi data latih dan data uji, pelatihan model, serta interpretasi hasil klasifikasi. Namun, penelitian ini tidak mencakup analisis mendalam terhadap faktor-faktor penyebab depresi, faktor geografis atau lokasi siswa, intervensi psikologis, maupun solusi klinis yang dapat diberikan. Selain itu, penelitian ini tidak bertujuan untuk memberikan diagnosis medis, melainkan hanya memanfaatkan pendekatan komputasi melalui teknik *machine learning* untuk mendukung proses deteksi dini depresi pada siswa.

Manfaat

Manfaat dari tugas akhir ini adalah:

1. Memberikan kontribusi dalam upaya deteksi dini masalah kesehatan mental di kalangan siswa melalui pendekatan berbasis data dan teknologi.
2. Menyediakan model prediktif yang dapat digunakan sebagai dasar dalam pengembangan sistem pendukung keputusan (*decision support system*) bagi lembaga pendidikan atau praktisi psikologi dalam memetakan siswa yang berisiko mengalami depresi.

3. Mendorong pemanfaatan *machine learning* di bidang kesehatan mental untuk meningkatkan kualitas intervensi dan penanganan secara lebih tepat sasaran dan preventif.

TINJAUAN PUSTAKA

Teknik Machine Learning untuk Prediksi Kesehatan Mental

Pada penelitian yang dilakukan oleh Jain *et al.* (2021) dalam jurnal IEEE yang menggunakan delapan model yaitu, *Decision Tree* (DT), *Random Forest* (RF), *Support Vector Machine* (SVM), *Naïve Bayes* (NB), *Logistic Regression* (LR), *XGBoos*t (XGB), *Gradient Boosting Classifier* (GBC), dan *Artificial Neural Network* (ANN) untuk memprediksi adanya indikasi depresi dengan 76 jenis atribut yang berbeda. Setelah menerapkan berbagai strategi dan model yang berbeda didapatkan hasil terbaik ada pada akurasi 87,38%, yang dicapai menggunakan metode *Support Vector Machine* (SVM).

Penerapan XGBoost untuk Deteksi Depresi

Dalam penelitian yang dilakukan oleh Sharma dan Verbeke (2020), dijelaskan bahwa diagnosis gangguan mental khususnya depresi hanya dilakukan melalui wawancara. Sehingga, penelitian ini mengembangkan model *Extreme Gradient Boosting* (XGBoost) untuk memprediksi adanya gangguan depresi. Tantangan yang dihadapi pada penelitian ini ada pada ketidakseimbangan kelas yaitu dengan 570 kasus depresi dari total 11.081, sehingga model cenderung bias terhadap kelas mayoritas (Tidak Depresi). Setelah dilakukan teknik resampling didapatkan akurasi model yang seimbang, precision, recall dan F1-score diatas 0,90.

Perbandingan Kernel Support Vector Machine (SVM)

Penelitian serupa juga dilakukan oleh Aulia *et al.* (2021) dalam Jurnal SINTECH Journal vol. 4 No 2, yang mengkaji terkait penerapan metode Support Vector Machine (SVM) pada analisis sentimen vaksinasi COVID-19 di Indonesia. Penelitian ini membandingkan empat jenis kernel SVM yaitu linear, sigmoid, polinomial dan radial basis function (RBF). Hasil analisis menunjukkan bahwa kernel linear dan sigmoid memberikan akurasi tertinggi, yaitu sebesar 87%. Sedangkan, kernel RBF dan polinomial menghasilkan akurasi sebesar 86%. Model dengan kernel linear dan sigmoid memprediksi jumlah negatif sebanyak 132, netral sebanyak 928 dan positif sebanyak 917. Penelitian memperkuat efektivitas SVM kernel sigmoid dalam melakukan klasifikasi.

Dari berbagai penelitian sebelumnya, diketahui bahwa SVM, XGBoost, dan Logistic Regression merupakan algoritma yang efektif untuk klasifikasi depresi. Masing-masing memiliki keunggulan, seperti mampu menangani data tidak seimbang atau kemudahan interpretasi. Oleh karena itu, penelitian ini dilakukan untuk membandingkan ketiga model tersebut dalam mendeteksi depresi pada siswa guna menemukan model terbaik.

METODE

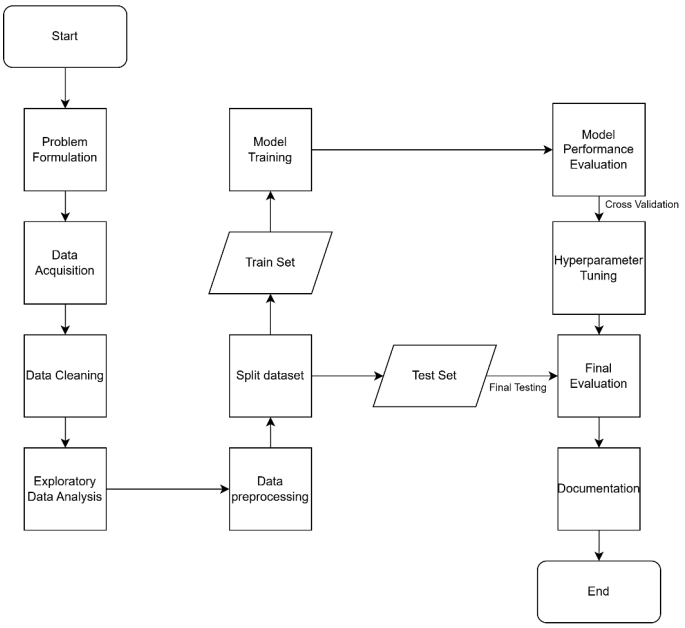
Data

Data yang digunakan pada penelitian ini dapat dilihat pada *link* berikut student-depression-dataset yang berasal dari Kaggle. Data tersebut berisikan 27.901 baris dengan 17 atribut, yaitu ID, Gender, Age, City, Profession, Academic Pressure, Work Pressure, CGPA, Study Satisfaction, Job Satisfaction, Sleep Duration, Dietary Habits, Degree, Have you ever had suicidal thoughts?, Work/Study Hours, Financial stress, dan

Family History of Mental Illness dengan variabel target Depression. Data tersebut merupakan data mentah sehingga tidak dapat langsung diproses untuk *data mining* karena terdapat *missing value* dan perlu melakukan seleksi terhadap beberapa atribut untuk meningkatkan akurasi model nantinya.

Tahapan Kegiatan

Tahapan kegiatan dalam projek analisis data depresi pada siswa tercantum dalam diagram alir yang tercantum pada Gambar 1.



Gambar 1 Diagram alir pengerjaan proyek

Langkah awal dalam proyek ini dimulai dengan perumusan masalah yang hendak diselesaikan, yaitu klasifikasi status depresi pada *student* yang meliputi siswa kelas 12 dan mahasiswa dari berbagai jenjang pendidikan. Tujuan utama dari proyek ini adalah mengembangkan sebuah model *machine learning* yang mampu mengklasifikasikan apakah seseorang mengalami depresi atau tidak, berdasarkan berbagai atribut yang tersedia dalam dataset. Permasalahan ini dipilih karena isu kesehatan mental, khususnya depresi pada pelajar dan mahasiswa, semakin mendapat perhatian luas dan membutuhkan penanganan berbasis data untuk mendukung deteksi dini.

Untuk menjawab permasalahan tersebut, kami menggunakan dataset yang bersumber dari platform Kaggle dengan judul Student Depression Dataset, yang dapat diakses melalui tautan student-depression-dataset. Dataset ini menyajikan berbagai fitur yang merepresentasikan faktor-faktor potensial yang memengaruhi kondisi depresi pada mahasiswa, seperti tekanan akademik, gaya hidup, serta riwayat kesehatan mental. Dengan karakteristik tersebut, dataset ini dinilai relevan untuk digunakan dalam membangun model klasifikasi depresi.

Setelah data dikumpulkan, proses pembersihan data (*data cleaning*) dilakukan guna memastikan bahwa data yang akan digunakan benar-benar bersih dan siap untuk dianalisis lebih lanjut. Beberapa tahapan pembersihan meliputi penanganan nilai hilang (*missing values*), penghapusan data duplikat, serta identifikasi dan penanganan *outlier*. Selain itu, transformasi data juga dilakukan agar format data menjadi seragam sehingga dapat diproses dengan lebih efektif pada tahap-tahap berikutnya.

Selanjutnya, proses *exploratory data analysis* (EDA) dilakukan untuk memahami lebih dalam karakteristik data. Analisis korelasi antarfitur dilakukan untuk mengidentifikasi fitur-fitur yang memiliki hubungan kuat dengan status depresi. Sebaran data untuk masing-masing kelas target (depresi dan tidak depresi) diamati untuk mengevaluasi potensi ketidakseimbangan kelas. Selain itu, distribusi data dan keberadaan *outlier* divisualisasikan guna memperoleh gambaran yang lebih komprehensif terkait kualitas data yang akan digunakan dalam pelatihan model.

Tahap terakhir dalam proses persiapan data adalah *preprocessing*, yang bertujuan untuk menyiapkan data agar sesuai dengan kebutuhan algoritma *machine learning*. Pada tahap ini, fitur-fitur kategorik diidentifikasi dan dikodekan (*encoding*) menjadi format numerik. Fitur-fitur yang tidak relevan terhadap prediksi target dihapus untuk mengurangi kompleksitas model. Selain itu, semua fitur numerik kemudian dilakukan normalisasi (*scaling*) agar berada pada skala yang seragam, sehingga model dapat belajar secara optimal.

Setelah melakukan *preprocessing*, dilakukan pembangunan model dengan menerapkan beberapa algoritma *machine learning*. Model-model dibangun dengan menggunakan bahasa pemrograman Python. Pembangunan model yang diawali dengan *data splitting.* Data dibagi menjadi dua bagian, antara lain *training set* yang digunakan untuk melatih model dan *test set* yang digunakan untuk evaluasi akhir model.

Algoritma *machine learning* yang dicoba pada penelitian ini antara lain Logistic Regression, SVM, dan XGBoost. Model-model tersebut dilatih menggunakan data dari *training set* agar dapat mempelajari pola-pola yang berhubungan dengan status depresi mahasiswa. Setelah proses pelatihan selesai, model kemudian diuji menggunakan metode *cross-validation* untuk mengetahui seberapa baik performanya sebelum dilakukan *tuning*. Metrik evaluasi yang digunakan meliputi *accuration*, *precision*, *recall*, dan *f1-score*. Lalu, agar model dapat mencapai performa terbaik, dilakukan proses *tuning* terhadap *hyperparameter*-nya. Dalam hal ini, kami menggunakan metode Randomized Grid Search untuk menjelajahi kombinasi *hyperparameter* secara efisien dan menemukan konfigurasi terbaik yang memberikan hasil optimal. Setelah mendapatkan model dengan *hyperparameter* terbaik, dilakukan evaluasi akhir menggunakan *test set*. Tujuannya adalah untuk mengukur kemampuan generalisasi model terhadap data yang benar-benar baru, dan memastikan bahwa performa yang baik tidak hanya terjadi pada data pelatihan (*train set*) saja. Setelah itu, dibuat *learning curve* pada masing-masing model untuk mengidentifikasi adanya indikasi *overfitting* atau *underfitting*. Setelah itu, dipilih model terbaik untuk diterapkan pada tahap selanjutnya.

Tahap terakhir pada penelitian ini adalah *documentation*. Tahapan ini mencakup penyusunan laporan akhir proyek secara terstruktur. Laporan dimulai dengan latar belakang masalah yang menjelaskan urgensi klasifikasi depresi pada mahasiswa, dilanjutkan dengan tinjauan pustaka yang membahas teori dan studi terdahulu terkait metode yang digunakan. Selanjutnya, bagian metodologi memaparkan langkah-langkah pengolahan data dan pelatihan model secara sistematis. Terakhir, disajikan hasil dan pembahasan yang berisi evaluasi performa model serta interpretasi dari temuan yang diperoleh selama proses analisis.

Lingkungan Pengembangan

Pengembangan dan analisis data dalam penelitian ini dilakukan menggunakan platform Google Collaboratory (Colab). Alasan utama pemilihan Google Colab adalah karena kemampuannya untuk menjalankan kode Python secara daring tanpa perlu instalasi lokal, serta kemudahan dalam berkolaborasi secara *real-time* dengan tim. *Platform* ini juga mendukung integrasi dengan Google Drive, memudahkan penyimpanan serta pembagian

file. Selain itu, untuk menunjang proses pengolahan dan analisis data, digunakan berbagai *library* Python yang umum dalam prosesnya, mulai dari eksplorasi data, pembersihan, pelatihan model, hingga evaluasi dapat dilakukan secara terstruktur dan efisien.

Library pandas digunakan untuk manipulasi dan analisis data berbentuk tabular, sementara numpy digunakan untuk operasi numerik yang efisien. Visualisasi data dilakukan dengan bantuan matplotlib.pyplot dan seaborn untuk mempermudah pembuatan grafik yang lebih informatif dan menarik secara tampilan.

Dalam tahap pemodelan, digunakan library scikit-learn (sklearn), yang menyediakan berbagai alat untuk *machine learning*, termasuk pembagian data menggunakan train_test_split, pelatihan model menggunakan algoritma seperti Support Vector Machine (SVM) melalui SVC, serta evaluasi model menggunakan metrik seperti akurasi, *confusion matrix*, dan *classification report*. Selain itu, digunakan juga teknik *cross-validation* (cross_val_score) dan pencarian parameter optimal menggunakan Grid Search (GridSearchCV) untuk meningkatkan performa model.

Agar data siap untuk diproses oleh model, digunakan *preprocessing* seperti normalisasi dan *encoding*. Dalam hal ini, digunakan StandardScaler untuk penskalaan data numerik agar berada pada skala yang seragam dan OrdinalEncoder untuk mengubah fitur kategorik menjadi bentuk numerik yang dapat dikenali oleh model pembelajaran mesin.

HASIL DAN PEMBAHASAN

Exploratory Data Analysis (EDA)

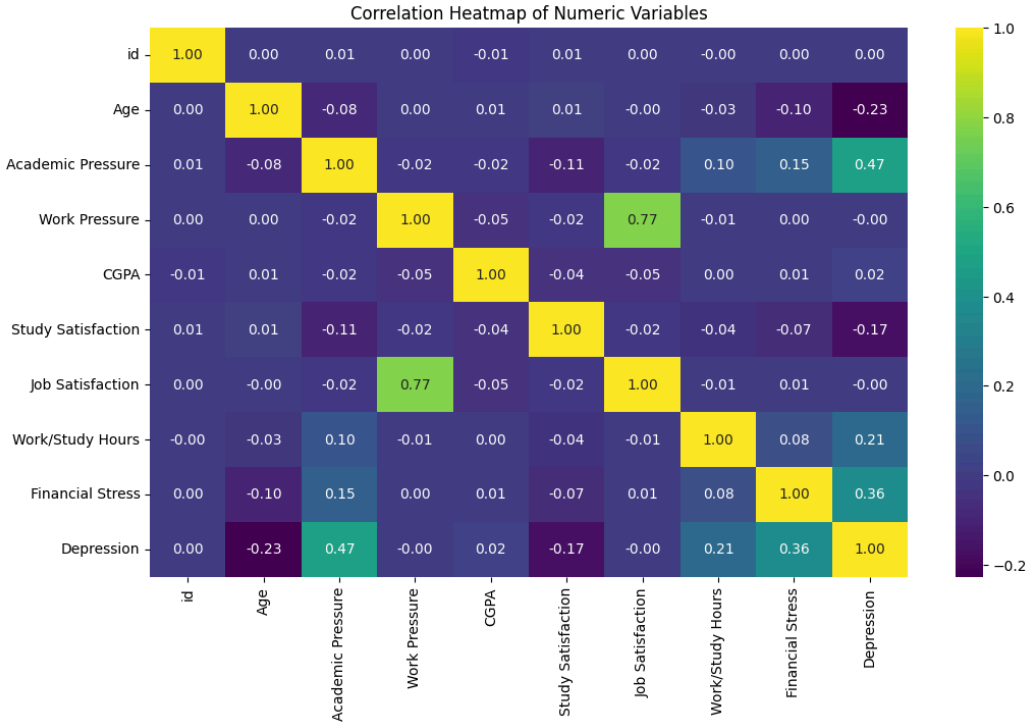
1. Statistik Deskriptif

	id	Age	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Work/Study Hours	Financial Stress	Depression
count	27901.000000	27901.000000	27901.000000	27901.000000	27901.000000	27901.000000	27901.000000	27901.000000	27898.000000	27901.000000
mean	70442.149421	25.822300	3.141214	0.000430	7.656104	2.943837	0.000681	7.156984	3.139867	0.585499
std	40641.175216	4.905687	1.381465	0.043992	1.470707	1.361148	0.044394	3.707642	1.437347	0.492645
min	2.000000	18.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
25%	35039.000000	21.000000	2.000000	0.000000	6.290000	2.000000	0.000000	4.000000	2.000000	0.000000
50%	70684.000000	25.000000	3.000000	0.000000	7.770000	3.000000	0.000000	8.000000	3.000000	1.000000
75%	105818.000000	30.000000	4.000000	0.000000	8.920000	4.000000	0.000000	10.000000	4.000000	1.000000
max	140699.000000	59.000000	5.000000	5.000000	10.000000	5.000000	4.000000	12.000000	5.000000	1.000000

Gambar 2 Statistik deskriptif

Berdasarkan hasil statistik deskriptif, data ini merepresentasikan populasi dengan usia rata-rata 25,8 tahun, yang kemungkinan besar adalah mahasiswa atau individu muda di awal karier. Tekanan akademik yang dirasakan berada pada tingkat sedang hingga tinggi (rata-rata 3,14 dari skala 5), sedangkan tekanan kerja memiliki median 0, yang mengindikasikan bahwa sebagian besar responden tidak bekerja atau tidak merasa tertekan oleh pekerjaan. Rata-rata nilai CGPA responden cukup tinggi, yaitu 7,66 dari maksimum 10, menunjukkan performa akademik yang relatif baik. Tingkat kepuasan terhadap studi berada pada rata-rata 2,94 (dari skala 5), sementara kepuasan terhadap pekerjaan cenderung rendah, yang selaras dengan asumsi bahwa sebagian besar belum bekerja secara penuh. Waktu kerja atau belajar per hari rata-rata adalah 7,16 jam, yang menunjukkan beban aktivitas harian cukup tinggi, dengan tekanan finansial juga berada di level sedang hingga tinggi (rata-rata 3,14 dari 5). Adapun distribusi kolom target juga menunjukkan nilai yang cukup seimbang dengan 58.5% responden mengalami depresi dan selebihnya tidak depresi. Secara keseluruhan, data ini memperlihatkan bahwa kelompok responden berada dalam tekanan akademik dan finansial yang signifikan, yang kemungkinan turut berkontribusi terhadap tingginya prevalensi depresi.

2. Korelasi Variabel Pada Dataset



Gambar 3 *Heatmap* korelasi pearson variabel

Dari *heatmap* di atas diidentifikasi fitur yang memiliki korelasi pearson yang cukup besar satu sama lain, dimana nilai korelasinya lebih besar 0.5. Variabel "Work Pressure" dan "Job Satisfaction" memiliki korelasi 0.7 yang menandakan ada indikasi redundansi pada kedua fitur ini. Oleh karena itu, perlu penanganan pada kedua fitur ini sebelum memasuki model untuk menghindari bias pada model yang berpotensi mengakibatkan overfitting.

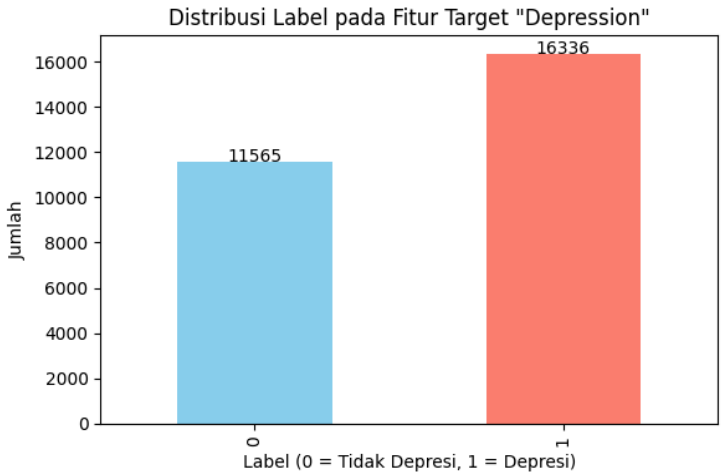
3. Multikolinearitas

```
===== [1] Cek Multikolinearitas (VIF) =====
                                    feature        VIF
0                                       Age  23.633688
1                          Academic Pressure   6.388804
2                                      CGPA  19.096902
3                        Study Satisfaction   5.273504
4                            Sleep Duration   2.487384
5                            Dietary Habits   2.292651
6                                    Degree   3.439348
7                          Work/Study Hours   4.675328
8                          Financial Stress   5.821761
9                               Gender_Male   2.242894
10  Have you ever had suicidal thoughts ?_Yes   3.062322
11       Family History of Mental Illness_Yes   1.922247
```

Gambar 4 Nilai VIF variabel

Hasil pemeriksaan multikolinearitas menggunakan Variance Inflation Factor (VIF) menunjukkan bahwa terdapat beberapa variabel dengan nilai VIF yang cukup tinggi, yang mengindikasikan adanya potensi multikolinearitas dalam data. Secara umum, nilai VIF di atas 10 dianggap menunjukkan multikolinearitas yang serius, sedangkan nilai antara 5 hingga 10 mengindikasikan tingkat yang moderat. Dalam output tersebut, variabel "Age" memiliki nilai VIF paling tinggi yaitu 23.63, diikuti oleh "CGPA" dengan VIF sebesar 19.10 dan "Academic Pressure" sebesar 6.39. Ketiga variabel ini patut dicermati karena dapat menyebabkan ketidakstabilan dalam estimasi koefisien model regresi, membuat interpretasi menjadi tidak akurat, dan meningkatkan varians parameter. Oleh karena itu, untuk mengatasi masalah multikolinearitas dapat digunakan model-model yang *robust* pada multikolinearitas.

4. Distribusi Kelas



Gambar 5 Distribusi kolom "Depression"

Berdasarkan visualisasi distribusi label pada fitur target "Depression", terlihat bahwa jumlah observasi antara kelas 0 (Tidak Depresi) dan kelas 1 (Depresi) tidak terlalu timpang. Kelas 0 berjumlah 11.565 dan kelas 1 berjumlah 16.336, dengan rasio kurang lebih 41% : 59%. Meskipun terdapat sedikit ketidakseimbangan, proporsi ini masih dalam batas wajar dan tidak tergolong sebagai kasus klasifikasi yang imbalance secara ekstrem, seperti yang umumnya terjadi jika rasio di bawah 20:80 atau bahkan 10:90. Oleh karena itu, penggunaan teknik penanganan imbalance seperti oversampling (SMOTE) atau undersampling tidak diperlukan dalam kasus ini, karena dapat justru menyebabkan overfitting atau hilangnya informasi penting dari data asli. Dengan distribusi target yang masih cukup seimbang ini, model klasifikasi seperti regresi logistik atau pohon keputusan dapat belajar secara efektif tanpa perlu intervensi tambahan untuk menyeimbangkan kelas. Sebaliknya, fokus sebaiknya diberikan pada pemilihan fitur yang tepat, validasi silang, dan evaluasi metrik yang sesuai (misalnya F1-score), agar model tetap adil dan akurat dalam mendeteksi kedua kelas.

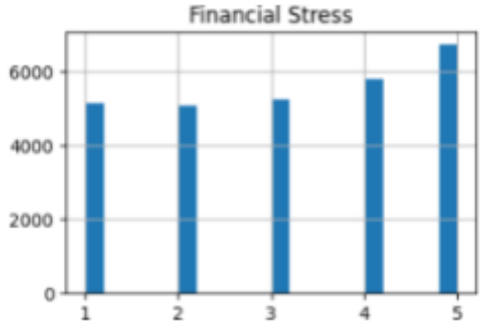
Data Preprocessing

Pada tahap *data preprocessing,* dilakukan serangkaian langkah untuk memastikan data yang digunakan memiliki kualitas yang baik dan cukup layak untuk dianalisis lebih

lanjut. Langkah-langkah ini mencakup penanganan nilai kosong, deteksi dan penanganan duplikasi serta *outlier*, seleksi fitur, dan *encoding* variabel kategorikal.

1. Penanganan Nilai Kosong

   Pada tahap awal pra-pemrosesan data, langkah pertama yang dilakukan adalah memastikan tidak adanya nilai kosong pada dataset. Proses pengecekan menunjukkan bahwa hanya kolom "Financial Stress" yang memiliki nilai kosong, yaitu sebanyak tiga entri. Kolom-kolom lainnya dinyatakan lengkap tanpa *missing value*. Penanganan nilai kosong pada kolom "Financial Stress" menggunakan metode imputasi dengan nilai modus. Nilai modus dipilih karena kolom ini bersifat kategorikal, tidak menunjukkan distribusi yang sangat menyimpang (*skewed*) seperti gambar 1, serta jumlah data kosong yang sangat kecil sehingga tidak mengganggu distribusi secara keseluruhan. Dengan demikian, ketiga nilai kosong tersebut diisi dengan nilai modus Financial Stress, yaitu 5.0.



Gambar 6 Distribusi kolom “Financial Stress”

2. Deteksi Duplikasi

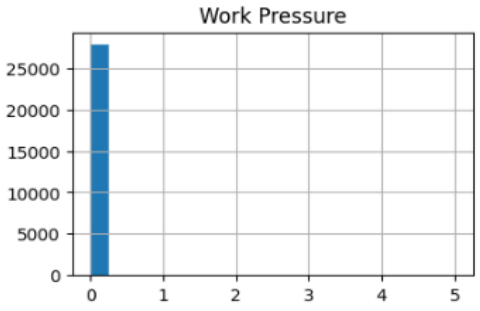
   Pengecekan duplikasi dilakukan untuk menghindari redundansi. Hasil analisis menunjukkan bahwa tidak terdapat satupun baris yang duplikat. Oleh karena itu, tidak dilakukan penghapusan data pada tahap ini.

3. Seleksi Fitur

   Langkah berikutnya adalah seleksi fitur untuk mengurangi fitur yang tidak relevan atau tidak informatif. Dari hasil observasi yang dilakukan ditemukan beberapa kolom yang tidak memberikan variasi atau informasi bermakna.

   a. Work Pressure

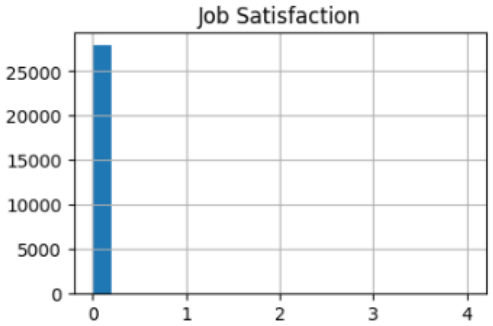
      Dapat dilihat dari gambar 2 bahwa kolom "Work Pressure" dieliminasi karena tidak memiliki variasi pada dataset. Melalui proses lebih lanjut, ditemukan hanya 3 entri dari 27870 entri yang memiliki nilai yang berbeda. Hampir seluruh observasi memiliki nilai “Work Pressure”sama yaitu di nilai 0. Hal tersebut berarti kolom ini tidak memberikan informasi, maka kolom ini diputuskan untuk dihapus.



Gambar 7 Distribusi kolom "Work Pressure "

b. Job Satisfaction

Serupa dengan kolom ″Work Pressure ″, kolom “Job Satisfaction” juga tidak memberikan variansi dan informasi pada analisis maupun model yang terlihat dari Gambar 3. Setelah diobservasi lebih jauh, ditemukan hanya 8 entri yang memiliki nilai bukan 0. Selain itu, kedelapan entri ini terdiri atas label 0 dan 1 yang terbagi rata, yaitu 4 entri berlabel 1 dan 4 entri berlabel 0. Oleh karena itu, disimpulkan kolom ini tidak memberikan informasi yang berpengaruh pada analisis maupun model nantinya.



Gambar 8 Distribusi kolom “Job Satisfaction”

c. Profession

Kolom “Profession” juga dihapus karena seluruh entri memberi informasi yang homogen. Kolom ini menyatakan profesi yang sama, yaitu "*student*", sehingga tidak menyediakan informasi pembeda.

d. Id

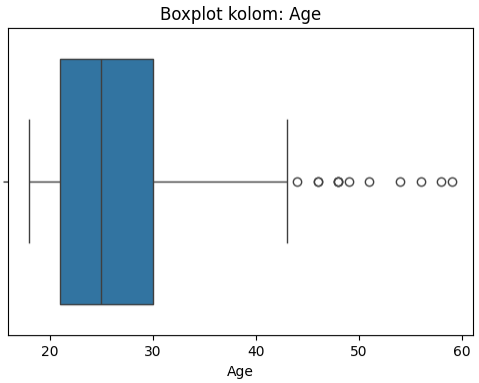
Berbeda dengan kolom “Profession”, kolom “Id” menampilkan nilai unik satu sama lain pada setiap observasi. Namun kolom ini tetap dihapus karena setiap baris memiliki ID unik yang hanya berfungsi sebagai penanda, bukan sebagai fitur yang dapat digunakan dalam analisis.

e. City

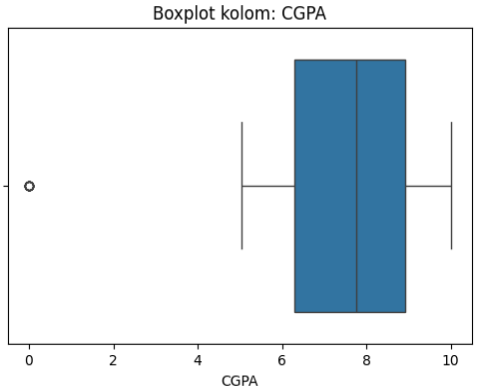
Penghapusan kolom “City” dilakukan karena lokasi geografis responden dianggap tidak relevan dalam konteks penelitian ini yang lebih menekankan pada kondisi psikologis individu, bukan aspek spasial.

4. Deteksi dan Penanganan Outlier

Langkah selanjutnya adalah mendeteksi dan menangani *outlier* dengan menggunakan metode Interquartile Range (IQR). Pada kolom “Age”, ditemukan sebanyak 12 titik yang dikategorikan sebagai *outlier*, terdiri dari 9 nilai unik. Namun, nilai-nilai tersebut tetap dipertahankan karena dianggap masih berada dalam batas kewajaran secara umum untuk usia mahasiswa. Di sisi lain, kolom “CGPA” menunjukkan adanya 9 outlier, namun semuanya memiliki nilai yang sama yaitu 0. Nilai “CGPA” sebesar 0 dianggap tidak wajar untuk seorang mahasiswa aktif, sehingga seluruh entri dengan nilai tersebut dihapus dari dataset agar tidak mempengaruhi hasil analisis secara negatif.



Gambar 9 *Outlier* kolom "Age" dengan metode IQR



Gambar 10 *Outlier* kolom "CGPA" dengan metode IQR

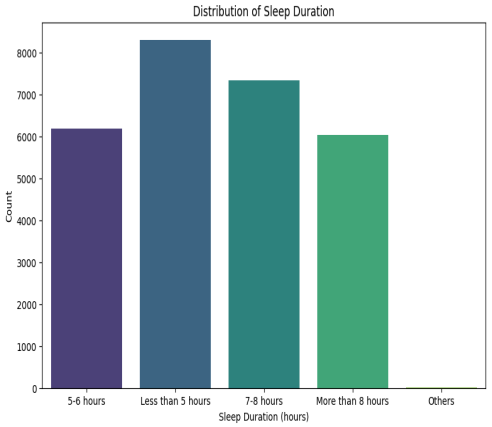
5. *Encod* Variabel Kategorikal

Setelah data dibersihkan dari nilai kosong, duplikasi, dan *outlier* yang tidak relevan, tahap terakhir pra-proses adalah *encoding* variabel kategorikal agar dapat digunakan dalam algoritma klasifikasi Support Vector Machine (SVM). Untuk menyiapkan data agar dapat digunakan dalam model SVM, dilakukan proses *encoding* terhadap variabel kategorikal, dengan dua metode:

a. *Ordinal Encoding*

i. Sleep Duration

Kolom ini terdiri dari 5 kategori, seperti yang ditampilkan pada gambar 6. Kategori diurutkan berdasarkan durasi tidur, di mana nilai *encoding* yang lebih tinggi menunjukkan durasi tidur yang lebih lama. *Mapping encoding* untuk variabel tersebut dapat dilihat pada tabel 1.



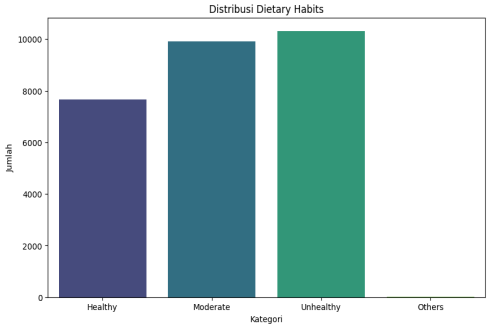
Gambar 11 Distribusi kolom "Sleep Duration"

Tabel 1 *Mapping encoding* kolom “Sleep Duration”

Kategori kolom “Sleep Duration”	**Ordinal encoding**
Others	0
Less than 5 hours	1
5-6 hours	2
7-8 hours	3
More than hours	4

ii. Dietary Habits

Kolom ini mencakup 4 kategori sebagaimana terlihat pada gambar. Nilai *encoding* yang lebih tinggi merepresentasikan kualitas pola makan yang lebih baik. Detail *mapping encoding* untuk variabel ini disajikan pada tabel 2.



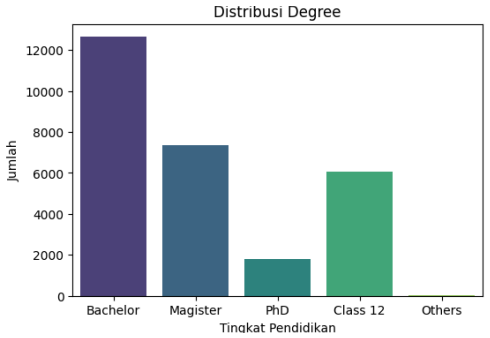
Gambar 8 Distribusi kolom “Dietary Habits”

Tabel 12 *Mapping encoding* kolom “Dietary Habits”

Kategori kolom “Dietary Habits”	**Ordinal encoding**
Others	0
Unhealthy	1
Moderate	2
Healthy	3

iii. Degree

Kolom *degree* di-*encode* secara ordinal berdasarkan jenjang pendidikan yang ada pada dataset ditunjukkan pada gambar 8, dimana nilai integer yang lebih tinggi diberikan pada tingkat pendidikan yang lebih tinggi. *Mapping* lengkap untuk encoding variabel ini disajikan pada tabel 3.



Gambar 13 Distribusi kolom "Degree"

Tabel 3 *Mapping encoding* kolom "Degree"

Kategori kolom "Degree"	**Ordinal encoding**
Others	0
Class 12	1
Bachelor	2
Magister	3

b. *One-Hot Encoding*

Metode *one-hot encoding* digunakan pada saat encoding kolom kategorikal nominal yang tidak memiliki urutan logis. Pada implementasinya digunakan fungsi *get_dummies* dari pustaka pandas. Hasil dari encoding ini adalah kolom-kolom baru seperti "Gender_Male", "Have you ever get suicidal thoughts?_Yes", *dan* Family "History of Mental Illness_Yes", yang masing-masing merepresentasikan kondisi biner dari fitur aslinya.

Berdasarkan seluruh tahap prapemrosesan, maka dapat disimpulkan fitur-fitur yang digunakan dalam pemodelan dapat dilihat pada Tabel 4.

Tabel 4 Prediktor yang digunakan untuk mendeteksi status depresi siswa

Prediktor	**Keterangan**
Age	Menyimpan informasi usia dari siswa. Nilainya berkisar dari 18 hingga 59 tahun dengan rata-rata 25.82
Academic Pressure	Menyimpan informasi tekanan akademik yang dialami siswa. Memiliki nilai dari skala 0 hingga 5 dengan 0 berarti "Academic Pressure" terendah sedangkan 5 berarti "Academic Pressure" paling tinggi
CGPA	Menyimpan informasi CGPA (*cumulative grade point average*) dari siswa. Nilainya merentang dari 0 hingga 10 dengan rata-rata 7.65

Study Satisfaction	Menyimpan informasi kepuasan belajar siswa. Memiliki nilai dari skala 0 hingga 5 dengan 0 berarti "Study Satisfaction" terendah sedangkan 5 berarti "Study Satisfaction" paling tinggi
Sleep Duration	Menyimpan informasi rata-rata jam tidur harian siswa. Kolom terdiri atas kategori "Others", "Less than 5 hours", "5-6 hours", "7-8 hours", "More than 8 hours". Hasil *ordinal encoding* secara berturut-turut adalah 0, 1, 2, 3, 4
Dietary Habits	Menyimpan informasi kebiasaan makan siswa. Kolom terdiri atas kategori "Others", "Unhealthy", "Moderate", "Healthy". Hasil *ordinal encoding* secara berturut-turut adalah 0, 1, 2, 3
Degree	Menyimpan informasi gelar dari siswa. Gelar yang terdapat dalam data, antara lain "B.Pharm", "BSc", "BA", "BCA", "M.Tech", "PhD", "Class 12", "B.Ed", "LLB", "BE", "M.Ed", "MSc", "BHM", "M.Pharm", "MCA", "MA", "B.Com", "MD", "MBA", "MBBS", "M.Com", "B.Arch", "LLM", "B.Tech", "BBA", "ME", "MHM", "Others". Untuk gelar yang diawali dengan huruf "B" di-*encode* sebagai 1, untuk gelar yang diawali dengan huruf "M" di-*encode* sebagai 2, dan selainnya di-*encode* sebagai "0"
Work/Study Hours	Menyimpan informasi rata-rata jam bekerja atau belajar siswa. Nilainya merentang dari 0 hingga 12 jam dengan rata-rata 7.16 jam
Financial Stress	Menyimpan informasi tingkat stres keuangan siswa. Memiliki skala nilai dari 1 hingga 5 dengan 1 adalah *stress* terendah sedangkan 5 adalah *stress* tertinggi
Gender_Male	Menyimpan informasi jenis kelamin siswa. Telah diterapkan *one hot encoding* pada prediktor ini, bernilai 1 jika "Male" dan bernilai 0 jika "Female"
Have you ever had suicidal thoughts ?_Yes	Menyimpan informasi apakah siswa pernah memiliki pikiran untuk bunuh diri. Telah diterapkan *one hot encoding* pada prediktor ini, bernilai 1 jika "Yes" dan bernilai 0 jika "No"

Family History of Mental Illness_Yes	Menyimpan informasi apakah terdapat riwayat gangguan mental pada keluarga. Telah diterapkan *one hot encoding* pada prediktor ini, bernilai 1 jika "Yes" dan bernilai 0 jika "No"

Pemodelan

Sebelum pemodelan dilakukan, kolom target dan kolom fitur dipisahkan. Kolom target, yakni "Depression", disimpan ke dalam variabel y, sementara itu kolom fitur disimpan ke dalam variabel X. Variabel X menyimpan kolom-kolom yang akan menjadi prediktor dari status depresi siswa yang telah diseleksi. Adapun prediktor-prediktor tersebut dapat dilihat pada Tabel 4. Kemudian, standarisasi diimplementasikan kepada variabel X menggunakan fungsi *standard scaler* dari *library* sklearn. Setelah itu, dataset dibagi menjadi dua bagian, yakni data latih dan data uji dengan rasio pembagian 80% untuk data latih dan 20% untuk data uji. Data dilatih menggunakan tiga jenis model klasifikasi, yakni model Logistic Regression, XGBoost, dan SVM (Support Vector Machine).

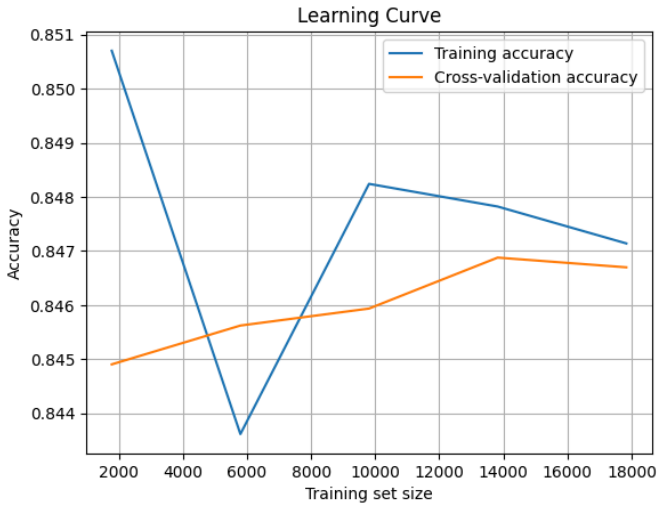
Dalam proses pelatihan model, digunakan teknik cross-validation untuk memperoleh hasil evaluasi yang lebih objektif dan *reliable*. Metode ini memungkinkan model untuk dilatih dan diuji secara bergantian pada seluruh bagian data, sehingga mengurangi risiko *overfitting* dan memberikan gambaran yang lebih menyeluruh terhadap performa model. Pada penelitian ini, diterapkan *k-fold cross-validation* dengan jumlah *fold* sebanyak lima, yang berarti seluruh data dibagi menjadi lima bagian yang relatif seimbang. Model kemudian dilatih pada empat bagian data dan divalidasi pada satu bagian sisanya secara bergantian hingga semua bagian mendapat giliran sebagai data validasi. Pendekatan ini membantu memastikan bahwa performa model tidak hanya bergantung pada subset tertentu dari data melainkan kemampuan generalisasi pada data secara keseluruhan.

Pemodelan Logistic Regression

Sebelum membangun model Logistic Regression, dilakukan proses *hyperparameter tuning* untuk memperoleh parameter yang menghasilkan performa terbaik. Proses *tuning* dilakukan menggunakan metode *grid search*, yang mengevaluasi berbagai nilai parameter C. Nilai-nilai C yang dipilih terdiri atas 0.01, 0.1, dan 1.0. Nilai-nilai tersebut dipilih karena parameter C mengatur kekuatan regularisasi, di mana nilai kecil mendorong model yang lebih sederhana untuk menghindari risiko terjadinya *overfitting*. Berdasarkan hasil *grid search*, diperoleh bahwa parameter yang paling optimal adalah C = 0.1.

Model Logistic Regression menunjukkan performa yang cukup baik, dengan akurasi pelatihan sebesar 84,68783% dan rata-rata akurasi validasi silang (*cross-validation*) sebesar 84,67434%. Untuk menggambarkan pergerakan performa model terhadap variasi ukuran data pelatihan, dibuat *learning curve* (kurva pembelajaran) sebagaimana ditampilkan pada Gambar 3. Kurva ini memperlihatkan perubahan akurasi pada data pelatihan dan data validasi silang seiring bertambahnya jumlah data yang digunakan dalam pelatihan, sehingga dapat memberikan gambaran mengenai kestabilan dan kemampuan generalisasi model. *Learning curve* menunjukkan bahwa pada awal pelatihan dengan ukuran *dataset* relatif sedikit, model menunjukkan akurasi pelatihan yang tinggi (sekitar 85%), namun akurasi validasinya masih rendah. Ini umum terjadi karena model cenderung *overfitting* terhadap data pelatihan yang jumlahnya sedikit. Saat ukuran data pelatihan

ditambah, akurasi pelatihan mengalami penurunan yang cukup signifikan, yakni ketika *dataset* berjumlah 6000, kemudian meningkat kembali. Ini mengindikasikan bahwa model sedang mengalami proses penyesuaian terhadap kompleksitas data yang semakin bertambah. Sementara itu, akurasi validasi cenderung meningkat secara bertahap dan stabil, menunjukkan perbaikan kemampuan generalisasi model seiring bertambahnya data. Menariknya, pada titik sekitar 10.000 data, akurasi validasi dan pelatihan menjadi lebih seimbang, dengan gap yang kecil. Hal ini merupakan indikator bahwa model tidak mengalami *overfitting* maupun *underfitting* yang signifikan, dan performa umumnya cukup stabil hingga ukuran data maksimum.



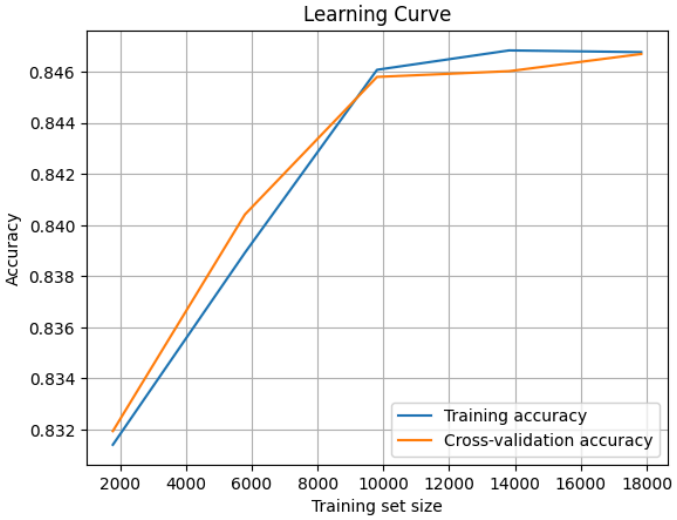
Gambar 14 *Learning curve* dari model Logistic Regression

Setelah model Logistic Regression diterapkan pada data *test*, diperoleh akurasi prediksi sebesar 84,66092%. *Confusion matrix* menunjukkan bahwa model berhasil mengklasifikasikan 1.841 data negatif (kelas 0/tidak depresi) dan 2.878 data positif (kelas 1/depresi) dengan benar, sementara terjadi kesalahan klasifikasi pada 506 data negatif dan 349 data positif. Berdasarkan *classification report*, model memperoleh *precision* 84%, *recall* 78%, dan *f1-score* 81% untuk kelas negatif. Ini berarti model cukup baik dalam memprediksi kelas negatif, meskipun masih ada proporsi yang salah diklasifikasikan sebagai positif. Sementara itu untuk kelas positif, model memiliki *precision* 85%, *recall* 89%, dan *f1-score* 87%, menunjukkan performa yang lebih stabil dalam mengidentifikasi kelas positif. Secara keseluruhan, rata-rata *f1-score* tertimbang (*weighted average*) adalah 85%, yang mengindikasikan bahwa model memberikan kinerja cukup *reliable* dalam mengklasifikasikan data. Hasil ini memperkuat temuan sebelumnya dari *learning curve*, yakni model memiliki kemampuan generalisasi yang baik pada data uji atau data yang belum pernah dilihat.

Pemodelan Support Vector Machine (SVM)

Sebelum membangun model Support Vector Machine (SVM), dilakukan proses *hyperparameter tuning* untuk memperoleh kombinasi parameter yang menghasilkan performa terbaik. Proses *tuning* dilakukan menggunakan metode *grid search*, yang mengevaluasi berbagai kombinasi nilai parameter C, *gamma*, dan jenis kernel. Rentang nilai dari parameter yang dipilih antara lain: C = {0.01, 0.1}; gamma = {0.01, 0.1}; dan kernel = {linear, poly, rbf, sigmoid}. Nilai-nilai C yang kecil dipilih karena akan memberikan regularisasi yang lebih kuat, sehingga cenderung menghasilkan model dengan

margin yang lebih lebar dan generalisasi yang lebih baik terhadap data baru. Sementara itu, nilai-nilai gamma yang kecil dipilih karena akan menghasilkan model yang lebih halus dengan pengaruh data yang tersebar lebih luas, sehingga model tidak sensitif terhadap data terdekat dari *hyperlane* yang dapat meningkatkan risiko *overfitting*. Adapun nilai-nilai kernel merupakan berbagai macam dari model SVM itu sendiri. Berdasarkan hasil *grid search*, diperoleh bahwa kombinasi parameter yang paling optimal adalah C = 0.01, *gamma* = 0.1, dan jenis kernel = 'sigmoid'. Model SVM dengan konfigurasi tersebut kemudian dilatih menggunakan data pelatihan, menghasilkan akurasi pelatihan sebesar 84,66989%. Evaluasi menggunakan validasi silang (*cross-validation*) menghasilkan rata-rata akurasi sebesar 84,66987%, yang menunjukkan konsistensi performa model serta kemampuan generalisasi yang baik terhadap data yang tidak terlihat sebelumnya. Gambar 11 menyajikan *learning curve* dari model SVM dengan kernel *sigmoid*. Dari grafik *learning curve*, dapat diamati bahwa peningkatan ukuran set pelatihan secara konsisten meningkatkan akurasi model, baik pada data pelatihan maupun validasi silang. Seiring bertambahnya jumlah data pelatihan menjadi 10.000, akurasi kedua kurva mengalami peningkatan signifikan dan mulai konvergen di nilai sekitar 0,846. Menariknya, setelah titik ini, yaitu pada ukuran data pelatihan 13.000 hingga 18.000, kurva akurasi menunjukkan kecenderungan melandai. Akurasi pelatihan sedikit lebih tinggi dibandingkan validasi silang, namun perbedaannya sangat kecil (< 0,001), yang menunjukkan bahwa model telah mencapai batas kemampuan generalisasinya terhadap data baru. Hal ini mengindikasikan bahwa penambahan data pelatihan lebih lanjut tidak memberikan peningkatan signifikan terhadap akurasi model. Dengan demikian, hasil *learning curve* ini menunjukkan bahwa model yang digunakan tidak mengalami *overfitting* maupun *underfitting* secara signifikan. Kedua kurva cenderung konvergen, yang merupakan indikasi bahwa model mampu belajar dengan baik dari data pelatihan dan memiliki kemampuan generalisasi yang memadai terhadap data yang tidak terlihat sebelumnya.

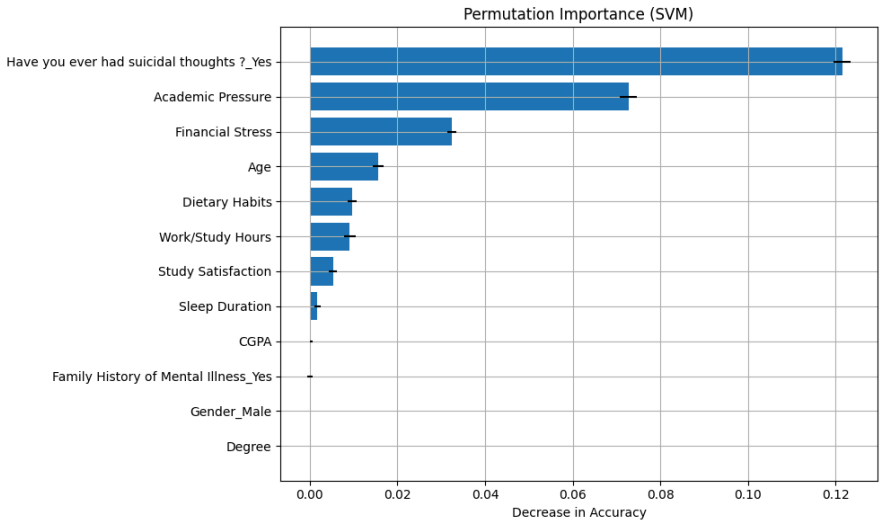


Gambar 15 *Learning curve* dari model Support Vector Machine (SVM)

Evaluasi performa model dilakukan menggunakan metrik akurasi, *confusion matrix*, dan *classification report* pada data pengujian. Model SVM menghasilkan tingkat akurasi keseluruhan sebesar 84,49946%. Berdasarkan *confusion matrix*, dari total 2.347 sampel kelas 0, sebanyak 1.807 diklasifikasikan dengan benar (*true negatives*), sementara 540 diklasifikasikan secara salah sebagai kelas 1 (*false positives*). Untuk kelas 1, dari 3.227 sampel, model berhasil mengklasifikasikan 2.903 sampel dengan benar (*true*

positives), sedangkan 324 sisanya salah diklasifikasikan sebagai kelas 0 (*false negatives*). Hasil *classification report* lebih lanjut menunjukkan bahwa model memiliki performa yang seimbang antar kelas. Untuk kelas 0, nilai *precision* sebesar 0.85 dan *recall* sebesar 0.77, menghasilkan *f1-score* 0.81. Sementara itu, untuk kelas 1, *precision* sebesar 0.84 dan *recall* 0.90, dengan *f1-score* 0.87. Rata-rata tertimbang (*weighted average*) dari *precision*, *recall*, dan *f1-score* masing-masing adalah 0.85, 0.84, dan 0.84, mencerminkan stabilitas kinerja model terhadap distribusi data yang tidak terlalu seimbang. Nilai *macro average* dari *f1-score* sebesar 0.84 mengindikasikan bahwa model memperlakukan kedua kelas secara relatif setara dalam hal performa klasifikasi. Tingginya nilai *recall* pada kelas 1 (0.90) menunjukkan bahwa model sangat efektif dalam mendeteksi kasus positif. Secara keseluruhan, hasil evaluasi menunjukkan bahwa model memiliki kinerja klasifikasi yang baik dan seimbang, dengan kesalahan klasifikasi yang relatif rendah serta tingkat generalisasi yang tinggi, sebagaimana juga tercermin dalam kurva pembelajaran sebelumnya.

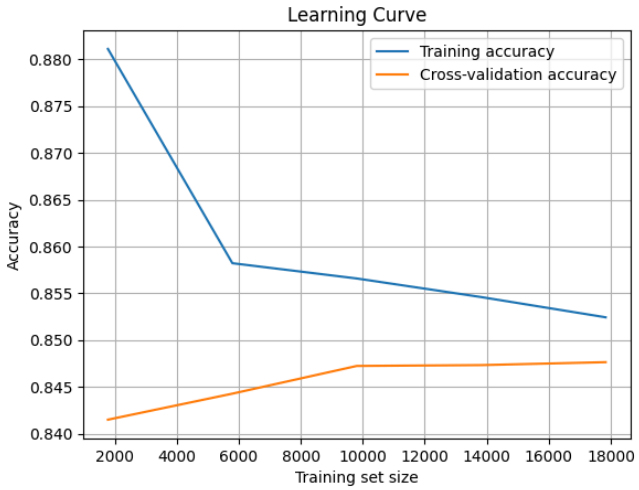
Selain mengevaluasi performa model secara kuantitatif, dilakukan pula analisis *feature importance* menggunakan pendekatan *permutation importance* terhadap model SVM. Hasil analisis *permutation importance* terhadap model Support Vector Machine (SVM) menunjukkan bahwa fitur "Have you ever had suicidal thoughts? _Yes" memiliki nilai kontribusi tertinggi terhadap performa model dalam memprediksi kondisi depresi, dengan rata-rata penurunan akurasi sebesar 0.1215 ketika fitur tersebut diacak. Hal ini mengindikasikan bahwa variabel ini merupakan prediktor yang sangat dominan, dan secara substansial memengaruhi *output* klasifikasi. Fitur "Academic Pressure" dan "Financial Stress" juga menunjukkan tingkat importance yang cukup tinggi, masing-masing dengan nilai 0.0727 dan 0.0324. Ini mencerminkan bahwa tekanan akademik dan stres keuangan merupakan determinan penting dalam munculnya gejala depresi, terutama dalam konteks populasi mahasiswa atau individu usia produktif. Sebaliknya, beberapa fitur menunjukkan nilai *importance* yang sangat rendah, seperti CGPA, Degree, Gender_Male, dan Family History of Mental Illness_Yes. Nilai-nilai yang mendekati nol ini mengimplikasikan bahwa keberadaan atau pengacakan fitur-fitur tersebut tidak menyebabkan penurunan akurasi prediksi yang berarti, sehingga relevansinya terhadap model dapat dipertanyakan. Dalam konteks pengembangan model prediktif yang efisien, fitur-fitur ini berpotensi untuk dieliminasi guna mengurangi kompleksitas model tanpa mengorbankan performa.



Gambar 16 *Feature importance* berdasarkan penurunan akurasi

Pemodelan XGBoost

Model terakhir yang digunakan adalah XGBoost. Sebelum proses pelatihan dilakukan, dilakukan *hyperparameter tuning* terlebih dahulu menggunakan metode *grid search* untuk memperoleh kombinasi parameter yang menghasilkan performa model paling optimal. Parameter yang disesuaikan meliputi *max_depth* yang berperan dalam mengontrol kompleksitas model, *learning_rate* yang menentukan laju pembelajaran model, dan *subsample* yang menentukan proporsi data yang akan digunakan dalam setiap iterasi pelatihan. Nilai-nilai dari parameter yang dipilih antara lain: *max_depth* = {2, 3, 5, 7}; *learning rate* = {0.1, 0.01, 0.001}; dan *subsample* = {0.2, 0.5, 0.7, 1}. Rentang nilai *max_depth* yang kecil dipilih untuk menghindari terjadinya *overfitting* dengan menghasilkan pohon yang dangkal. Sementara itu, nilai-nilai *learning rate* yang kecil membuat proses pelatihan lebih lambat namun stabil dan akurat, sehingga menghindari terjadinya *overfitting.* Adapun nilai-nilai pada *subsample* bertujuan mengeksplorasi sejauh mana penggunaan sebagian data dapat meningkatkan generalisasi model. Hasil *grid search* menunjukkan bahwa kombinasi parameter terbaik untuk model ini adalah *max_depth* = 3, *learning_rate* = 0.1, dan *subsample* = 0.5. Setelah pelatihan dengan konfigurasi tersebut dilakukan, diperoleh akurasi pada data *training* sebesar 85,2081% dan akurasi *cross-validation* yang dirata-ratakan adalah sebesar 84,78647%. *Learning curve* dari model XGBoost yang tersaji dalam Gambar 13 menunjukkan bahwa akurasi pelatihan mengalami penurunan seiring dengan bertambahnya jumlah data pelatihan, dari sekitar 88,1% pada ukuran data 2.000 menjadi sekitar 85,2% pada ukuran data maksimum. Penurunan ini merupakan indikasi bahwa model menjadi lebih general dan tidak terlalu menghafal data latih (*overfitting*), seiring meningkatnya ukuran sampel pelatihan. Sebaliknya, akurasi validasi silang menunjukkan tren peningkatan yang konsisten, dari sekitar 84,1% hingga mendekati 84,8%. Hal ini mencerminkan peningkatan kemampuan generalisasi model terhadap data yang tidak terlihat, seiring dengan bertambahnya informasi dari data pelatihan. Selain itu, jarak antara kurva pelatihan dan validasi secara bertahap mengecil, yang merupakan indikasi bahwa model mulai mencapai keseimbangan antara bias dan *variance*. Stabilitas kurva validasi pada titik-titik akhir juga menunjukkan bahwa penambahan data pelatihan tidak lagi memberikan peningkatan akurasi yang signifikan, yang dapat diartikan bahwa model telah konvergen dengan konfigurasi *hyperparameter* yang digunakan.



Gambar 17 *Learning curve* dari model XGBoost

Evaluasi akhir terhadap model XGBoost dilakukan menggunakan data uji yang tidak pernah dilibatkan dalam proses pelatihan maupun validasi. Hasil pengujian menunjukkan bahwa model berhasil mencapai tingkat akurasi sebesar 84,76856%. Analisis lebih lanjut melalui *confusion matrix* menunjukkan bahwa model mampu mengklasifikasikan 1.844 sampel kelas negatif (label 0) dan 2.881 sampel kelas positif (label 1) secara benar, sementara terjadi kesalahan klasifikasi pada 503 sampel kelas negatif dan 346 sampel kelas positif. Pola distribusi ini menunjukkan bahwa model memiliki kecenderungan yang relatif seimbang dalam menangani kedua kelas. Berdasarkan *classification report*, diperoleh nilai *precision* sebesar 0.84 untuk kelas negatif dan 0.85 untuk kelas positif, yang menunjukkan tingkat ketepatan prediksi model untuk masing-masing kelas. Sementara itu, nilai *recall* menunjukkan bahwa model mampu mengidentifikasi 79% dari total sampel kelas negatif dan 89% dari sampel kelas positif dengan benar. Nilai *f1-score* yang dihasilkan mencapai 0.81 untuk kelas negatif dan 0.87 untuk kelas positif. Nilai rata-rata tertimbang (*weighted average*) untuk *precision*, *recall*, dan *f1-score* masing-masing sebesar 0.85, 0.85, dan 0.85 yang menunjukkan bahwa model XGBoost memiliki performa yang seimbang dalam melakukan klasifikasi pada kedua kelas target.

Perbandingan Model

Di antara ketiga model yang dilatih, yaitu Logistic Regression, Support Vector Machine (SVM), dan XGBoost, model SVM dipilih sebagai model utama karena menunjukkan performa klasifikasi yang paling stabil dan konsisten. Meskipun secara akurasi model ini sedikit lebih rendah dibandingkan XGBoost (84,49946% vs. 84,76856%) dan Logistic Regression (84,49946% vs. 84,66092%), namun kestabilan performanya yang tercermin dari nilai akurasi pelatihan dan validasi silang yang hampir identik (masing-masing sebesar 84,66989% dan 84,66987%) serta bentuk kurva pembelajarannya yang konvergen menjadi pertimbangan utama dalam pemilihan model. Kurva pembelajaran SVM menunjukkan peningkatan yang stabil dan mendekati titik konvergensi, dengan gap antara akurasi pelatihan dan validasi yang sangat kecil (< 0,001) pada ukuran data pelatihan maksimum. Ini menandakan bahwa model tidak mengalami *overfitting* maupun *underfitting* yang signifikan serta telah mencapai kapasitas belajarnya secara optimal. Selain itu, model SVM menunjukkan performa yang seimbang dalam mengklasifikasikan kedua kelas target, dengan *precision* dan *recall* yang tinggi dan hampir setara pada masing-masing kelas. Nilai *f1-score* untuk kelas negatif dan positif masing-masing sebesar 0.81 dan 0.87, serta rata-rata tertimbang sebesar 0.85, menegaskan bahwa model mampu melakukan prediksi secara andal dan adil terhadap distribusi kelas yang tidak seimbang. Selain itu, performa *recall* pada model ini paling tinggi dibandingkan dengan model lainnya, yakni mencapai 90%. Nilai *recall* yang tinggi menunjukkan kemampuan model dalam mendeteksi sebanyak mungkin kasus positif, dalam hal ini individu yang berisiko mengalami gangguan kesehatan mental. Dalam konteks kesehatan mental, tingkat kesalahan berupa *false negative*—yaitu kasus positif yang tidak terdeteksi—dapat berimplikasi serius, karena individu yang sebenarnya membutuhkan perhatian justru luput dari perhatian. Oleh karena itu, prioritas utama dalam pemodelan ini adalah meminimalkan kesalahan tersebut, sehingga pemilihan model dengan recall tertinggi menjadi keputusan strategis untuk mendukung deteksi dini yang lebih efektif dan akurat.

Sebaliknya, meskipun XGBoost menunjukkan performa akhir yang sedikit lebih tinggi dari segi akurasi dan *recall*, model ini memperlihatkan dinamika kurva pembelajaran yang fluktuatif pada awal proses pelatihan dan baru menunjukkan kestabilan

setelah ukuran data cukup besar, yang bisa menjadi indikasi kebutuhan data lebih banyak untuk mencapai stabilitas yang sama dengan SVM. Logistic Regression pun menunjukkan kestabilan yang serupa dengan SVM dalam hal generalisasi model, namun kurva pembelajarannya mengindikasikan fluktuasi yang lebih tajam pada akurasi pelatihan di awal sebelum mencapai keseimbangan, serta nilai *f1-score* pada kelas negatif yang sedikit lebih rendah. Berdasarkan keseluruhan hasil evaluasi, model SVM dipilih karena menawarkan *trade-off* terbaik antara akurasi, stabilitas performa, dan kemampuan generalisasi, menjadikannya solusi yang lebih dapat diandalkan untuk implementasi pada data nyata.

KESIMPULAN DAN SARAN

Berdasarkan hasil evaluasi terhadap ketiga model yang dilatih, yaitu Logistic Regression, Support Vector Machine (SVM), dan XGBoost, model SVM dipilih sebagai model utama karena menunjukkan performa klasifikasi yang paling stabil dan konsisten. Meskipun tingkat akurasi SVM (84,49946%) sedikit lebih rendah dibandingkan XGBoost (84,76856%) dan Logistic Regression (84,66092%), kestabilan performa yang ditunjukkan melalui nilai akurasi pelatihan dan validasi silang yang hampir identik (masing-masing sebesar 84,66989% dan 84,66987%) serta bentuk kurva pembelajaran yang konvergen menjadi dasar utama dalam pemilihan model. Kurva pembelajaran SVM memperlihatkan peningkatan yang stabil dengan selisih akurasi pelatihan dan validasi yang sangat kecil pada ukuran data maksimum, yang mengindikasikan bahwa model tidak mengalami *overfitting* maupun *underfitting* secara signifikan. Selain itu, model SVM juga menunjukkan kinerja yang seimbang dalam mengklasifikasikan kedua kelas target, sebagaimana tercermin dari nilai *f1-score* yang tinggi dan relatif setara untuk kelas negatif (0,81) dan positif (0,87), serta nilai rata-rata tertimbang sebesar 0,85. Model ini juga dipilih karena memiliki nilai *recall* tertinggi sebesar 90%, yang menunjukkan kemampuan optimal dalam mendeteksi individu dengan risiko gangguan kesehatan mental. Dalam konteks deteksi medis, nilai recall yang tinggi memiliki peran krusial dalam meminimalkan terjadinya *false negative*, yaitu kondisi ketika individu yang sebenarnya mengidap penyakit (kasus positif) terklasifikasi secara keliru sebagai tidak sakit (kasus negatif). Situasi ini sangat berisiko karena dapat menyebabkan keterlambatan penanganan. Oleh karena itu, peningkatan recall secara langsung mendukung efektivitas deteksi dini dan memastikan bahwa intervensi medis dapat diberikan secara lebih tepat sasaran. Berdasarkan keseluruhan hasil evaluasi, SVM dipandang memberikan *trade-off* yang optimal antara akurasi, kestabilan performa, dan kemampuan generalisasi, sehingga dianggap paling layak untuk diimplementasikan pada data nyata.

Optimalisasi lebih lanjut pada model SVM dapat difokuskan pada aspek lain, seperti *feature engineering*, *hyperparameter tuning*, atau arsitektur model, daripada sekadar menambah ukuran data pelatihan, mengingat *learning curve* menunjukkan bahwa model telah mencapai titik konvergensi dan tambahan data tidak memberikan peningkatan akurasi yang signifikan. Dengan pendekatan optimalisasi yang lebih terarah, diharapkan performa model dapat ditingkatkan secara signifikan baik dari sisi akurasi keseluruhan maupun keandalan dalam mengklasifikasikan kelas.

DAFTAR PUSTAKA

Aulia TMP, Arifin N, Mayasari R. 2021. Perbandingan kernel support vector machine (svm) dalam penerapan analisis sentimen vaksinasi covid-19. SINTECH Journal. 4(2): 139-145. doi:10.31598/sintechjournal.v4i2.773.

Chung J, Teo J. 2022. Mental health prediction using machine learning: taxonomy, applications, and challenges. *Applied Computational Intelligence and Soft Computing*. 2022(1): 1–19. doi:10.1155/2022/9970363.

Garriga R, Mas J, Abraha S, Nolan J, Harrison O, Tadros G, Matic, A. 2022. Machine learning model to predict mental health crises from electronic health records. *Nature medicine*. 28(6): 1240-1248. doi:10.1038/s41591-022-01811-5.

Glaz AL, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, *et al.*. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5): 15708. doi:10.2196/15708.

Horwitz LI, Kuznetsova M, Jones SA. 2019. Creating a learning health system through rapid-cycle, randomized testing. *New England Journal of Medicine*. 381(12): 1175–1179. doi:10.1056/nejmsb1900856.

Iyortsuun NK, Kim S, Jhon M, Yang H, Pant S. 2023. Review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare*. 11(3): 285. doi:10.3390/healthcare11030285.

Jain T, Jain A, Hada PS, Kumar H, Verma VK, Patni A. 2021. Machine learning techniques for prediction of mental health. Proceedings Of The Third International Conference On Inventive Research In Computing Applications (ICIRCA); 2021 Sept 02-04; Coimbatore, India. Coimbatore: IEEE. hlm 1606-1613.

Paton F, Wright K, Ayre N. 2016. Improving outcomes for people in mental health crisis: a rapid synthesis of the evidence for available models of care. *Health Technologyl Assessment*. 20(3): 1-162. doi:10.3310/hta20030.

Sharma S, Verbeke M. 2020. Improving diagnosis of depression with xgboost machine learning model and a large biomarkers dutch dataset (n = 11,081). *Frontiers in Big Data*. 3:15. doi:10.3389/fdata.2020.00015

[WHO] World Health Organization. 2024. Mental Health of Adolescents.