ANALISIS REGRESI UNTUK FAKTOR-FAKTOR YANG MEMPENGARUHI TINGKAT STRES PADA REMAJA

Makalah ini disusun untuk memenuhi Project Mata Kuliah Model Linear.

Dosen Pengampu: Madona Yunita Wijaya, M.Sc

Disusun Oleh:

Nazwah Laeza Camelia	11230940000034
Annisa Kusuma Wardani	11230940000038
Syahrul Mauhub Yasser	11230940000054
Siti Aminah	11230940000068

PROGRAM STUDI MATEMATIKA FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH

JAKARTA

2025M/1446 H

ABSTRAK

Penelitian ini menganalisis faktor-faktor yang memengaruhi tingkat stres pada remaja menggunakan metode analisis regresi linear berganda. Data bersumber dari dataset publik 'Mental Health Analysis among Teenagers' yang berisi 5.000 entri dan 11 variabel, dengan *Wearable_Stress_Score* sebagai variabel respon. Variabel prediktor meliputi usia, jenis kelamin, durasi tidur, durasi olahraga, waktu penggunaan layar, durasi penggunaan media sosial, sistem dukungan sosial, dan performa akademik. Penelitian ini bertujuan mengidentifikasi faktor signifikan, mengukur besar pengaruh prediktor terhadap: *Wearable_Stress_Score*, serta mengevaluasi performa model regresi berdasarkan R^2, Adjusted R^2, RMSE, dan signifikansi statistik. Metode yang digunakan adalah *Ordinary Least Squares* (OLS) untuk meminimalkan selisih kuadrat antara nilai aktual dan prediksi, serta dilakukan uji signifikansi koefisien regresi. Proses *Exploratory Data Analysis* (EDA) dilakukan untuk memahami struktur data, mendeteksi *missing values* dan duplikasi, serta menganalisis distribusi dan korelasi antar variabel. Seleksi model menggunakan pendekatan *Stepwise Regression* (forward selection, backward elimination, dan both) serta *Best Subsets Regression* untuk mendapatkan model yang paling efisien dengan mempertimbangkan *Adjusted* R^2, MSE, AIC, BIC, dan PRESS. Uji asumsi regresi (multikolinearitas, normalitas residual, homoskedastisitas, dan independensi) dilakukan untuk memastikan validitas model. Evaluasi dan validasi model menggunakan metrik seperti R-squared, Adjusted R-squared, RMSE, MSE, dan uji signifikansi koefisien (Uji-t dan p-value), serta metode K-Fold Cross Validation untuk memastikan kemampuan generalisasi model.

Kata Kunci: Regresi Linear Berganda, Tingkat Stres Remaja, Wearable Stress Score, Seleksi Model, Uji Asumsi Regresi.

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh

Alhamdulillah puji syukur kita panjatkan kepada Allah SWT., yang telah memberikan Kesehatan dan rahmat-Nya sehingga kami dapat menyelesaikan Laporan Projek Akhir Mata Kuliah Model Linear yang berjudul **"Analisis Regresi Untuk Faktor-Faktor Yang Mempengaruhi Tingkat Stres Pada Remaja"**.

Shalawat serta salam tak lupa kami haturkan kepada Baginda Nabi Muhammad SAW. beserta kepada keluarga dan para sahabatnya. Laporan ini merupakan tugas akhir semester 4 mata kuliah Model Linear Program Studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta.

Kami menyadari bahwa dalam menyelesaikan Laporan Projek Akhir tidak terlepas dari dukungan dan bantuan dari berbagai pihak. Oleh karena itu, kami sangat berterima kasih kepada semua pihak yang telah membantu dalam penulisan laporan ini dan memohon maaf apabila masih banyak kekurangan dalam penulisan laporan ini. Kami mengharapkan kritik dan saran yang membangun dari pembaca agar kedepannya kami dapat memperbaiki dan mendapatkan hasil yang lebih baik. Selain itu, kami juga berharap semoga laporan ini dapat bermanfaat bagi kami dan pihak lain yang membacanya.

Wassalammu'alaikum Warahmatullahi Wabarakatuh

Jakarta, 06 Juli 2025

Penulis

DAFTAR ISI

BAB I

PENDAHULUAN

1.1. Latar Belakang

Kesehatan mental merupakan aspek penting yang mempengaruhi kualitas hidup seseorang secara menyeluruh, baik dari segi fisik, psikologis, maupun sosial. Dalam kehidupan sehari-hari, individu sering kali mengalami tekanan dari berbagai sumber seperti tuntutan akademik, pekerjaan, lingkungan sosial, dan penggunaan teknologi digital yang semakin intens. Hal ini dapat memicu tingkat stres yang tinggi dan berdampak negatif terhadap kesehatan mental dan fisik seseorang.

Seiring dengan perkembangan teknologi, memungkinkan pemantauan kondisi kesehatan mental secara lebih objektif melalui perangkat wearable yang mampu merekam indikator fisiologis dan menghasilkan wearable stress score yaitu skor yang merepresentasikan tingkat stres berdasarkan data biometrik pengguna. Adanya skor ini membuka peluang untuk menganalisis faktor-faktor yang berperan dalam mempengaruhi tingkat stres pada remaja secara kuantitatif dan terukur.

Berbagai faktor diduga memiliki kontribusi terhadap peningkatan atau penurunan tingkat stres, seperti usia, jenis kelamin, durasi tidur, kebiasaan berolahraga, lama waktu penggunaan media sosial, waktu yang dihabiskan di depan layar, dukungan sosial, serta performa akademik. Namun, belum semua faktor tersebut diketahui seberapa besar pengaruhnya terhadap tingkat stres yang terukur secara fisiologis melalui *wearable stress score*, sehingga perlu dilakukan kajian yang lebih mendalam.

Penelitian ini bertujuan untuk mengidentifikasi factor-faktor yang berpengaruh secara signifikan terhadap *wearable stress score* pada remaja menggunakan metode analisis regresi linear berganda. Metode ini digunakan untuk meneliti hubungan antara satu variabel terikat (respon) dengan beberapa variabel bebas (prediktor) sekaligus, sehingga dapat diketahui faktor-faktor mana saja yang memiliki pengaruh signifikan terhadap skor stress pada wearable stress score. Dengan hasil analisis ini, diharapkan dapat diperoleh pemahaman yang lebih mendalam mengenai penyebab stres serta menjadi dasar untuk upaya pencegahan dan penanganan stres secara lebih efektif, khususnya pada kalangan remaja

1.2. Rumusan Masalah

1. Faktor-faktor apa saja yang secara signifikan mempengaruhi tingkat stres pada remaja yang diukur melalui *wearable stress score* sebagai variabel respon?
2. Seberapa besar pengaruh masing-masing variabel prediktor terhadap variabel respon yaitu *wearable stress score*?
3. Bagaimana performa model regresi yang dibangun dalam memprediksi tingkat stress pada remaja yang diukur melalui variabel respon *wearable stress score,* berdasarkan nilai evaluasi seperti adj-R^2, RMSE, dan signifikansi statistik masing-masing variabel prediktor?

1.3. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk:

1. Untuk mengidentifikasi faktor-faktor yang secara signifikan mempengaruhi tingkat stres pada remaja, sebagaimana diukur melalui *wearable stress score* sebagai variabel respon.
2. Untuk mengetahui mengukur seberapa besar pengaruh masing-masing variabel prediktor yang signifikan terhadap *wearable stress score* menggunakan analisis regresi linear berganda.
3. Untuk membangun dan mengevaluasi performa model regresi linear dalam memprediksi *wearable stress score*, berdasarkan metrik evaluasi seperti R^2, Adjusted R^2, RMSE, dan signifikansi statistik masing-masing variabel prediktor.

1.4. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat, baik secara akademis maupun praktis. Penelitian ini diharapkan dapat memberikan kontribusi dalam bidang statistik, khususnya dalam penerapan analisis regresi linear berganda untuk mengidentifikasi dan mengukur pengaruh faktor-faktor yang secara signifikan mempengaruhi tingkat stres pada remaja, sebagaimana diukur melalui variabel respon *wearable stress score*. Selain itu, penelitian ini juga dapat menjadi referensi bagi penelitian selanjutnya yang ingin mengkaji topik serupa dengan pendekatan kuantitatif berbasis data biometrik. Hasil penelitian ini diharapkan dapat memberikan gambaran mengenai model prediktif yang dapat dimanfaatkan untuk memahami penyebab stres pada remaja secara lebih objektif dan terukur, serta mendukung upaya pencegahan dan penanganan stres yang berbasis data.

BAB II

METODE PENELITIAN

2.1. Deskripsi *Dataset*

Data yang digunakan dalam penelitian ini bersumber dari dataset publik berjudul Mental Health Analysis among Teenagers yang tersedia di platform Kaggle. Dataset ini berisi sebanyak 5.000 entri dengan 11 variabel yaitu User_ID (identitas unik responden), Age (usia responden), Gender (jenis kelamin), Sleep_Hours (rata-rata jam tidur per hari), Exercise_Hours (durasi olahraga harian), Screen_Time_Hours (yang dihabiskan di depan layer), dan Social_Media_Hours (durasi penggunaan media social), Support_System (tingkat dukungan sosial), Academic_Performance (performa akademik responden), serta Survei_Stress_Score (skor stres berdasarkan persepsi pribadi atau hasil kuesioner), dan Wearable_Stress_Score (skor stres fisiologis berdasarkan data biometrik yang diperoleh dari perangkat wearable) yang berisi informasi mengenai kondisi psikologis remaja yang diperoleh melalui survei serta hasil pemantauan perangkat *wearable*.

Wearable_Stress_Score berperan sebagai variabel tak bebas atau variabel respon (juga dikenal sebagai variabel dependent, outcome, atau target), karena nilainya akan diprediksi menggunakan model regresi. Variabel ini bersifat kontinu dan mencerminkan tingkat stres objektif responden. Sementara itu, variabel lainnya digunakan sebagai variabel bebas (independent, explanatory, atau input) yang diduga memiliki pengaruh terhadap tingkat stres. Variabel bebas dalam dataset ini terdiri dari dua jenis, yaitu kontinu dan kategorik. Variabel kontinu mencakup Age, Sleep_Hours, Exercise_Hours, Screen_Time_Hours, dan Social_Media_Hours, Seluruh variabel ini dipertimbangkan dalam analisis karena secara teoritis memiliki hubungan yang relevan dengan kondisi stres pada remaja. Analisis regresi linear berganda digunakan dalam penelitian ini untuk mengidentifikasi dan mengukur sejauh mana variabel-variabel tersebut berpengaruh secara signifikan terhadap skor stres yang diukur secara fisiologis.

2.2. Teknik Regresi Linear

Penelitian ini menggunakan pendekatan regresi linear berganda untuk menganalisis hubungan antara satu variabel respon kontinu, yaitu Wearable_Stress_Score, dengan sejumlah variabel prediktor sekaligus. Tujuan utama dari pemodelan ini adalah untuk mengidentifikasi faktor-faktor yang secara statistik

berpengaruh signifikan terhadap tingkat stres pada remaja, serta mengukur seberapa besar kontribusi masing-masing variabel prediktor terhadap perubahan skor stres tersebut.

Secara umum, model regresi linear berganda dapat dituliskan dalam bentuk persamaan matematis sebagai berikut:

$$Yi = \beta 0 + \beta 1X1i + \beta 2X2i + \cdots + \beta pXpi + \epsilon i \ , \ i = 1,2, \dots , n$$

- Yi adalah nilai variabel *Wearable_Stress_Score* (variabel respon) untuk pengamatan ke-i.
- Xji adalah adalah nilai dari variabel prediktor ke-j untuk pengamatan ke-i, j= 1,2, … , p.
- $\beta 0$ adalah intercept (konstanta) model
- $\beta j, j = 1,2, \dots , p$ adalah koefisien regresi yang berasosiasi dengan variabel prediktor ke- j
- ϵi adalah komponen eror (residual) untuk pengamatan ke-i yang diasumsikan berdistribusi normal dengan mean 0 dan variansi σ^2 , dan ϵi saling independent.

Model ini dibangun menggunakan metode *Ordinary Least Squares (OLS)*, yaitu pendekatan yang meminimalkan jumlah kuadrat selisih antara nilai aktual dan nilai prediksi dari variabel respon. Seluruh variabel prediktor dalam model diperiksa secara simultan untuk menentukan kontribusinya terhadap variabel respon, baik dari sisi arah hubungan (positif/negatif) maupun besar pengaruhnya. Uji signifikansi dilakukan terhadap masing-masing koefisien regresi untuk mengidentifikasi faktor-faktor yang berpengaruh nyata secara statistik terhadap tingkat stres fisiologis pada remaja.

Selain itu, model ini juga akan dievaluasi berdasarkan beberapa metode seperti R^2, Adjusted R^2, RMSE, dan MSE, serta diuji terhadap asumsi-asumsi regresi untuk memastikan validitas hasil yang diperoleh.

2.3. Exploratory Data Analysis

Tahapan *Exploratory Data Analysis* (EDA) dilakukan untuk memahami karakteristik umum data sebelum membangun model regresi. Langkah ini bertujuan untuk mengeksplorasi distribusi variabel, mendeteksi adanya nilai ekstrim (*outlier*), menangani data yang hilang (*missing values*), serta mengidentifikasi pola-pola awal dan hubungan antar variabel.

Langkah-langkah EDA dalam penelitian ini meliputi:

a. Pemeriksaan Struktur Data

Pemeriksaan ini mencakup jenis data dari setiap variabel (numerik/kategorik), banyaknya entri (baris), dan nama-nama kolom. Hal ini penting untuk memastikan data terbaca dengan baik dan sesuai untuk analisis lanjutan.

b. Identifikasi Data Hilang dan Duplikasi

   Dilakukan pengecekan apakah terdapat nilai yang kosong (*missing values*) atau baris data yang duplikat. Jika ditemukan, data hilang akan diproses lebih lanjut sesuai dengan jumlah dan dampaknya terhadap analisis, sedangkan duplikasi akan dihapus.

c. Statistika Deskriptif

   Statistik seperti nilai minimum, maksimum, mean, median, dan standar deviasi dihitung untuk variabel numerik seperti Age, Sleep_Hours, Exercise_Hours, Screen_Time_Hours, Social_Media_Hours, Wearable_Stress_Score, dan Survei_Stress_Score. Langkah ini memberikan gambaran umum mengenai sebaran dan kecenderungan nilai data.

d. Distribusi dan Visualisasi Data

   Distribusi variabel numerik divisualisasikan dengan histogram dan boxplot untuk melihat bentuk sebaran data serta mendeteksi kemungkinan outlier. Sementara itu, variabel kategorik seperti Gender, Support_System, dan Academic_Performance divisualisasikan dengan bar chart untuk melihat proporsi tiap kategori.

e. Deteksi Outlier

   Boxplot digunakan untuk mengidentifikasi nilai ekstrim yang mungkin mempengaruhi hasil analisis. Keberadaan outlier dievaluasi lebih lanjut untuk diputuskan apakah akan dikeluarkan, dipertahankan, atau ditransformasi.

f. Analisis Korelasi

   Dilakukan analisis korelasi Pearson antar variabel numerik untuk melihat kekuatan dan arah hubungan antara variabel-variabel prediktor dengan variabel respon. Hasil korelasi divisualisasikan dalam bentuk heatmap.

g. Analisis Hubungan Awal antar Variabel

   Scatter plot digunakan untuk melihat pola hubungan antara wearable stress score dengan masing-masing variabel prediktor numerik. Selain itu, *grouped boxplot* digunakan untuk mengeksplorasi perbedaan rata-rata wearable stress score berdasarkan kategori pada variabel kategorik.

Hasil EDA menjadi dasar dalam menentukan perlakuan terhadap data sebelum masuk ke tahap pemodelan, seperti transformasi data, penghapusan outlier, serta seleksi awal variabel yang relevan.

2.4. Seleksi Model

Setelah dilakukan eksplorasi data dan analisis awal, tahap selanjutnya dalam penelitian ini adalah membangun model regresi linear berganda yang bertujuan untuk mengidentifikasi dan mengukur pengaruh masing-masing variabel prediktor terhadap tingkat stres pada remaja, sebagaimana diukur melalui wearable stress score. Dalam proses pembangunan model, digunakan dua pendekatan utama, yaitu pembangunan model penuh (full model) dan seleksi model (model parsimonious).

Model penuh (full model) merupakan model awal yang dibangun dengan memasukkan seluruh variabel prediktor yang tersedia dalam dataset tanpa melakukan seleksi terlebih dahulu. Model ini bertujuan untuk melihat pengaruh kolektif dari semua variabel terhadap variabel respon dan memberikan gambaran menyeluruh atas struktur data. Full model sering kali digunakan sebagai titik awal dalam analisis regresi, serta sebagai dasar pembanding terhadap model-model hasil seleksi, baik dari segi kelengkapan informasi maupun performa prediksi.

Namun, memasukkan semua variabel ke dalam model tidak selalu menghasilkan model yang efisien. Beberapa variabel mungkin tidak berkontribusi secara signifikan terhadap variabel respon atau bahkan dapat menimbulkan masalah multikolinearitas dan overfitting, yang dapat menurunkan validitas model. Oleh karena itu, diperlukan proses seleksi variabel untuk membangun model yang lebih sederhana

Dalam penelitian ini, metode seleksi yang dipertimbangkan antara lain:

a. Stepwise Regression: Metode ini melakukan penambahan dan/atau penghapusan variabel secara bertahap berdasarkan nilai signifikansi statistik (p-value) dan kontribusinya terhadap peningkatan *Adjusted* R^2. Proses stepwise dilakukan dalam dua arah, yaitu forward selection dan backward elimination, dan both untuk menemukan kombinasi variabel yang optimal.
    - Forward selection adalah metode yang memulai dari model kosong (hanya intercept) dan menambahkan variabel satu per satu. Pada setiap langkah, algoritma memilih variabel yang paling signifikan secara statistik (misalnya dengan p-value terkecil) dan yang paling meningkatkan performa model, seperti Adjusted R^2 atau menurunkan AIC. Proses ini terus berlanjut hingga tidak ada lagi variabel yang layak ditambahkan (yaitu, yang lolos ambang batas signifikansi).
    - Backward elimination adalah metode yang memulai dari model lengkap (semua variabel dimasukkan) dan secara bertahap menghapus variabel yang tidak signifikan. Pada setiap langkah, variabel dengan p-value tertinggi (biasanya > 0.05) dihapus dari model. Proses berhenti ketika semua variabel yang tersisa signifikan, atau ketika penghapusan tidak lagi meningkatkan model.
    - Metode ini menggabungkan forward dan backward. Proses dimulai dari model kosong atau sederhana, lalu menambahkan variabel signifikan seperti forward selection. setelah penambahan, diperiksa apakah variabel yang sudah ada masih

signifikan — jika tidak, maka dihapus seperti backward elimination. proses ini terus berlangsung dua arah (penambahan dan penghapusan) hingga model stabil.

b. Best Subsets Regression: Teknik ini mengevaluasi semua kemungkinan subset dari variabel prediktor dan membandingkan model-model yang dihasilkan berdasarkan kriteria statistik tertentu, seperti *Adjusted R²*, MSE, Mallow's Cp atau *AIC* (Akaike Information Criterion), BIC (Bayesian Information Criterion)). Metode ini membantu mengidentifikasi model dengan performa terbaik tanpa harus mengandalkan prosedur iteratif seperti pada stepwise.

- *Adjusted R²* (*Adjusted Coefficient of Determination*) → (0 sampai 1) : Mengukur proporsi varians yang dijelaskan oleh model, disesuaikan dengan jumlah variabel, mencegah bias terhadap model dengan banyak prediktor, semakin tinggi Adjusted R^2, semakin baik modelnya.
- MSE (*Mean Squared Error*) -> kurang dari 0 : Rata-rata kuadrat dari selisih antara nilai observasi dan nilai prediksi. Digunakan untuk menilai ketepatan model. Semakin kecil MSE, semakin baik model tersebut dalam memprediksi data.
- AIC (*Akaike Information Criterion)* dan BIC (*Bayesian Information Criterion*) : Digunakan untuk membandingkan model berdasarkan fit dan penalti kompleksitas. AIC lebih permisif; BIC lebih ketat terhadap banyaknya variabel. Semakin kecil nilai AIC/BIC, semakin baik modelnya.
- PRESS (*Prediction Sum of Squares*) : Ukuran validasi silang (cross-validation) leave-one-out. Mengukur seberapa baik model memprediksi data yang tidak digunakan dalam fitting. Semakin kecil PRESS, semakin baik performa prediksi model.

Hasil dari seleksi model akan menentukan variabel-variabel mana yang dimasukkan ke dalam model regresi linear berganda final. Model terpilih kemudian dievaluasi berdasarkan uji asumsi dan evaluasi model untuk memastikan validitas dan performanya.

2.5. Uji Asumsi Regresi

a. Uji Multikolinearitas

Uji Multikolinearitas merupakan hubungan linier antara variabel independen di dalam regresi berganda. Multikolinearitas terjadi ketika dua atau lebih variabel prediktor dalam sebuah model regresi memiliki korelasi yang tinggi, sehingga menyebabkan ketidakstabilan estimasi koefisien regresi. Salah satu cara untuk mendeteksi keberadaan multikolinearitas adalah dengan menggunakan Variance Inflation Factor (VIF) untuk setiap variabel prediktor. Nilai VIF yang melebihi ambang batas umum (biasanya diatas 5 atau 10) mengindikasikan adanya multikolinearitas yang signifikan. Jika multikolinearitas ditemukan, maka dapat dipertimbangkan untuk menghapus, menggabungkan, atau melakukan transformasi terhadap variabel terkait. Model regresi yang baik seharusnya tidak saling

berkorelasi secara signifikan antar variabel independen. Berikut adalah indikasi multikolinearitas:

- Korelasi tinggi antara dua variabel prediktor.
- Uji F signifikan, tetapi uji t untuk masing-masing variabel tidak signifikan. Ini menunjukkan bahwa secara keseluruhan model bagus, tapi tiap variabel tidak jelas kontribusinya.
- Estimasi koefisien regresi dengan tanda yang berlawanan dari yang diharapkan.
- Nilai VIF (Variance Inflation Factor) > 10 untuk parameter β, karena dianggap menunjukkan multikolinearitas serius yang kemungkinan besar memengaruhi estimasi koefisien secara signifikan (bisa dipertimbangkan untuk menghapus, menggabungkan, atau melakukan transformasi terhadap variabel terkait), dimana

$$(VIF)_i = \frac{1}{1 - R_i^2}, i = 1,2,\ldots,k$$

dan R_i^2 adalah koefisien determinasi dari model.

$$E(x_i) = \alpha_o + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + \cdots + \alpha_k x_k$$

b. Uji Normalitas

Asumsi normalitas residual ini menyatakan bahwa residual dari model regresi harus terdistribusi normal. Uji ini penting untuk menjamin validitas dari pengujian signifikansi koefisien regresi, terutama saat akan membuat kesimpulan dari data. Uji normalitas dapat dilakukan secara visual melalui histogram atau Q-Q plot, atau secara statistik dengan uji Shapiro-Wilk atau Kolmogorov–Smirnov. Jika pengujian menunjukkan penyimpangan dari distribusi normal, dapat dilakukan transformasi terhadap variabel atau menggunakan metode estimasi yang robust terhadap pelanggaran normalitas.

- Histogram (Grafik batang yang menampilkan distribusi nilai residual dari model regresi) : Untuk melihat apakah bentuk distribusi residual menyerupai distribusi normal (yaitu seperti kurva lonceng/simetris). Interprestasinya adalah Distribusi normal dimana histogram berbentuk simetris dan lonceng (bell-shaped). Penyimpangan dari normalitas terlihat berdasarkan kemiringan kurvan ke kiri/kanan (skewed), memiliki puncak tinggi (leptokurtic) atau tersebar rata (platykurtic), terlihat outlier atau nilai ekstrem.
- Q-Q plot (Grafik yang membandingkan kuantil residual aktual dengan kuantil dari distribusi normal teoretis.) : untuk Menguji apakah nilai residual mengikuti pola distribusi normal. Interpretasinya Titik-titik data mengikuti garis lurus diagonal yang berarti residual mendekati normal. Titik-titik melengkung ke atas atau bawah berarti penyimpangan dari normalitas. Melengkung ke atas/bawah di ujung berarti ada outlier atau distribusi heavy-tailed. S bentuk S atau Z yang berarti ada indikasi skewness (kemiringan).
- Uji Statistik Normalitas :
  a. Shapiro-Wilk Test : Uji statistik yang mengevaluasi kesesuaian distribusi data dengan distribusi normal. Umumnya digunakan untuk sampel kecil hingga menengah (n < 5000). Hipotesis nol (H_0) berarti data berasal dari

distribusi normal. Hipotesis alternatif (H_1) berarti data tidak berasal dari distribusi normal. Interprestasinya adalah jika p-value > 0.05 maka artinya gagal menolak H_0, artinya data tidak berbeda signifikan dari distribusi normal, maka normalitas terpenuhi. Jika p-value < 0.05 artinya menolak H_0, berarti data menyimpang dari normalitas.

b. Kolmogorov–Smirnov (K–S) Test : Uji statistik yang membandingkan distribusi kumulatif residual dengan distribusi normal standar. Cocok untuk sampel besar. Lebih umum digunakan untuk uji normalitas umum, tapi kurang sensitif terhadap perbedaan pada ekor distribusi dibanding Shapiro-Wilk. Interpretasinya adalah Jika p-value > 0.05 artinya distribusi tidak berbeda signifikan dari normal, maka normalitas terpenuhi. Jika p-value < 0.05 artinya data menyimpang dari distribusi normal, normalitas tidak terpenuhi

Solusi untuk pelanggaran asumsi normalitas diantaranya adalah:

- jika distribusi residual menunjukkan skew positif, dapat dilakukan transformasi terhadap variabel respon Y, misalnya menggunakan fungsi logaritmik log(Y) atau $\sqrt{Y}$
- Jika distribusi residual menunjukkan skew negatif, transformasi yang dapat dilakukan antara lain $\frac{1}{Y}$ atau $\frac{1}{\sqrt{Y}}$
- Alternatif lain, gunakan transformasi Box-Cox untuk secara otomatis memilih transformasi terbaik. Fungsi transformasi Box-Cox adalah:

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$

dimana parameter λ diestimasi menggunakan metode maksimum likelihood.

Catatan:

- Pelanggaran normalitas residual yang tidak parah umumnya dapat ditoleransi, asal ukuran sampel cukup besar (n ≥ 30 atau lebih).
- Transformasi variabel Y, seperti log(Y) atau Box-Cox, tidak hanya membantu mengatasi ketidaknormalan, tetapi juga bisa mengatasi heteroskedastisitas.
- Transformasi Box-cox tersedia juga di package MASS::boxcox(), caret::BoxCoxTrans().
- Fungsi Box-Cox hanya berlaku jika semua nilai Y > 0.Jika terdapat nilai nol atau negatif, data harus terlebih dahulu ditambah konstanta positif agar seluruh nilai menjadi positif sebelum transformasi diterapkan.

c. Uji Homoskedastisitas

Asumsi homoskedastisitas menyatakan bahwa residual memiliki variansi yang konstan di seluruh rentang nilai prediksi. Pelanggaran terhadap asumsi ini disebut heteroskedastisitas, yang dapat mengakibatkan estimasi varians koefisien menjadi kurang akurat atau tidak dapat diandalkan, sehingga mempengaruhi hasil uji statistik. Untuk mendeteksi kondisi ini, dilakukan analisis terhadap plot residual vs nilai prediksi serta pengujian statistik seperti Uji Breusch-Pagan atau Uji White. Jika

ditemukan gejala heteroskedastisitas atau pelanggaran asumsi homoskedastisitas, langkah yang dapat dilakukan adalah:

- Ketika asumsi homoskedastisitas tidak terpenuhi, bisa menggunakan transformasi Box-cox untuk menstabilkan variansi.
- Alternatifnya, menggunakan solusi weighted least square (WLS) untuk mengestimasi koefisien regresi. WLS memberikan bobot berbeda ke setiap observasi. Observasi dengan varian besar diberi bobot kecil, dan sebaliknya. Tujuannya agar residual terstandarisasi dan model lebih akurat.

$$min \sum_{i=1}^{n} \quad w_i(y_i - \hat{y}_i)^2$$

dimana $wi = 1/\sigma i$ 2 adalah bobot untuk observasi ke-i

- Bisa juga melakukan transformasi data atau menggunakan metode regresi robust.

d. Uji Independent

Asumsi independen menyatakan bahwa residual antar pengamatan harus bersifat independen (tidak berkorelasi). Jika residual saling berhubungan ini disebut autokorelasi, dan dapat membuat estimasi ketidakpastian hasil menjadi kurang akurat, yang melemahkan keandalan hasil statistik. Untuk mendeteksi adanya autokorelasi, digunakan uji Durbin-Watson. Nilai statistik Durbin-Watson yang mendekati 2 mengindikasikan bahwa tidak ada autokorelasi. Meskipun asumsi ini lebih relevan untuk data runtun waktu, tetap perlu dipertimbangkan dalam konteks data observasional terutama jika data berasal dari urutan logis atau waktu pengisian survei. Jika terdapat pelanggaran asumsi independent, langkah yang harus dilakukan adalah

- Jika data cross-section, yaitu data diambil secara acak dan tidak berkelompok, asumsi independent umumnya otomatis terpenuhi.
- Jika data mengandung unsur waktu (data time series), maka gunakan metode seperti ARIMA.
- Jika data spasial (wilayah berdekatan), bisa menggunakan metode seperti SAR.
- Jika data mengandung unsur kelompok (clustered), bisa menggunakan metode seperti mixed models, GEE maka asumsimengandung unsur waktu

2.6. Evaluasi dan Validasi Model

Setelah model regresi linear dibangun dan diuji asumsi-asumsi dasarnya, tahap selanjutnya adalah melakukan evaluasi terhadap performa model serta validasinya untuk memastikan bahwa model yang dihasilkan tidak hanya sesuai dengan data pelatihan (training data), tetapi juga mampu melakukan generalisasi dengan baik terhadap data baru.

a. Evaluasi Model

Evaluasi performa model dilakukan dengan mengukur sejauh mana model mampu menjelaskan variasi dari variabel respon (wearable stress score) serta seberapa akurat model memprediksi nilai output. Beberapa metrik evaluasi yang digunakan antara lain:

- R-squared (R^2): Menunjukkan proporsi variasi dari variabel respon yang dapat dijelaskan oleh variabel prediktor dalam model. Nilai R^2 berkisar antara 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan model yang lebih baik.
- Adjusted R-squared: Merupakan versi penyesuaian dari R^2 yang memperhitungkan jumlah variabel dalam model. Metrik ini lebih tepat digunakan untuk model dengan banyak prediktor, karena penalti diberikan untuk variabel yang tidak memberikan kontribusi signifikan.
- Root Mean Squared Error (RMSE): Mengukur rata-rata kesalahan prediksi dan nilai aktual, dalam satuan yang sama dengan variabel respon. Nilai RMSE yang lebih kecil mengindikasikan model yang lebih akurat dalam melakukan prediksi.
- Mean Squared Error (MSE): Menghitung rata-rata dari kuadrat selisih antara nilai aktual dengan nilai prediksi. Metrik ini memberikan penalti yang lebih besar terhadap kesalahan prediksi yang ekstrem karena kesalahan dikuadratkan. Oleh karena itu, MSE lebih sensitif terhadap outlier dan sering digunakan ketika kesalahan besar perlu ditekan. Nilai MSE yang lebih kecil menunjukkan performa model yang lebih baik secara keseluruhan.
- Signifikansi Statistik Koefisien (Uji-t dan p-value): Untuk menguji apakah setiap variabel prediktor secara statistik memiliki pengaruh signifikan terhadap wearable stress score. Variabel dengan p-value < 0.05 biasanya dianggap signifikan.

b. Validasi Model

Validasi model bertujuan untuk mengevaluasi kemampuan generalisasi model dalam melakukan prediksi yang akurat terhadap data baru (unseen data). Dalam penelitian ini, validasi dilakukan dengan metode K-Fold Cross Validation, yaitu data dibagi menjadi *k* bagian (fold), dan proses pelatihan serta pengujian dilakukan sebanyak *k* kali, setiap kali menggunakan fold yang berbeda sebagai data uji.

Rata-rata dari skor evaluasi setiap fold (seperti RMSE, MSE, atau R^2) menjadi indikator validitas model secara umum. Validasi silang ini membantu menghindari overfitting karena model diuji pada banyak bagian data yang berbeda dan memastikan bahwa model tidak hanya cocok terhadap data pelatihan, tetapi juga

dapat menggeneralisasi dengan baik. Validasi model juga bertujuan untuk memastikan stabilitas hasil estimasi, jika model menunjukkan performa yang stabil di semua fold, maka model dapat dianggap reliable.

BAB III

HASIL DAN PEMBAHASAN

3.1 Exploratory Data Analysis

3.1.1 Cek Struktur Data

Langkah pertama dalam proses analisis data adalah memahami keseluruhan struktur dari dataset yang akan digunakan. Hal ini bertujuan untuk memastikan bahwa data yang tersedia telah sesuai, lengkap, dan siap untuk dianalisis lebih lanjut. Oleh karena itu, dilakukan eksplorasi awal terhadap data melalui beberapa tahapan, baik dari segi bentuk, dimensi, nama-nama kolomnya, jenis data yang dimiliki setiap kolom, hingga ringkasan statistik dari variabel-variabel yang ada. Setiap tahapan ini disertai dengan potongan kode program serta output-nya yang ditampilkan dalam bentuk gambar, yang mendukung proses analisis menjadi jelas dan terbuka untuk dilihat.

1. Menampilkan 5 baris pertama dan 5 baris terakhir

   Langkah pertama yang dilakukan adalah menampilkan beberapa baris teratas dari dataset untuk mendapatkan gambaran umum mengenai isi data. Hal ini penting dilakukan untuk mengecek apakah data berhasil dibaca dengan benar oleh sistem, serta untuk mengamati struktur dan format penulisan data (misalnya apakah angka desimal sudah terdeteksi, apakah ada karakter yang tidak lazim, dan sebagainya).

```
df = pd.read_csv('mental_health_analysis.csv', encoding='latin-1')
df
```

	User_ID	Age	Gender	Social_Media_Hours	Exercise_Hours	Sleep_Hours	Screen_Time_Hours	Survey_Stress_Score	Wearable_Stress_Score	Support_System	Academ
0	1	16	F	9.654486	2.458001	5.198926	8.158189	3	0.288962	Moderate	
1	2	17	M	9.158143	0.392095	8.866097	5.151993	5	0.409446	Moderate	
2	3	15	M	5.028755	0.520119	4.943095	9.209325	2	0.423837	Moderate	
3	4	17	F	7.951103	1.022630	5.262773	9.823658	5	0.666021	Moderate	
4	5	17	F	1.357459	1.225462	6.196080	11.338990	5	0.928060	High	
...	...	...	...	...	...	...	...	...	...	...	
4995	4996	14	M	0.088148	1.003339	8.684888	5.922202	1	0.750205	Moderate	
4996	4997	15	F	7.161276	1.024644	5.312684	10.224924	4	0.427209	Moderate	
4997	4998	14	M	3.444383	2.877972	9.227726	4.059322	4	0.002893	High	
4998	4999	18	F	7.866525	2.395839	4.317831	10.657076	2	0.612063	High	
4999	5000	18	M	3.389362	1.375646	8.693171	6.977589	5	0.952662	Moderate	

5000 rows × 11 columns

2. Penghapusan variabel User_ID

   Dalam dataset ini terdapat kolom User_ID yang berfungsi sebagai nomor identifikasi unik untuk setiap responden. Kolom ini bersifat administratif dan

tidak mengandung informasi yang bersifat analitis maupun relevan terhadap prediksi skor stres atau variabel lainnya.

Karena User_ID tidak memiliki nilai informatif dalam analisis statistik, maka kolom tersebut dihapus sebelum proses eksplorasi data dimulai. Penghapusan ini dilakukan untuk menjaga kebersihan data dan menghindari penggunaan variabel yang tidak memberikan kontribusi terhadap interpretasi maupun pemodelan.

```
# Menghapus index berupa "User ID"
df.drop('User_ID', axis=1, inplace=True)
```

```
df
```

	Age	Gender	Social_Media_Hours	Exercise_Hours	Sleep_Hours	Screen_Time_Hours	Survey_Stress_Score	Wearable_Stress_Score	Support_System	Academic_Perfor
0	16	F	9.654486	2.458001	5.198926	8.158189	3	0.288962	Moderate	Exc
1	17	M	9.158143	0.392095	8.866097	5.151993	5	0.409446	Moderate	
2	15	M	5.028755	0.520119	4.943095	9.209325	2	0.423837	Moderate	
3	17	F	7.951103	1.022630	5.262773	9.823658	5	0.666021	Moderate	Av
4	17	F	1.357459	1.225462	6.196080	11.338990	5	0.928060	High	
...	...	...	...	...	...	...	...	...	...	
4995	14	M	0.088148	1.003339	8.684888	5.922202	1	0.750205	Moderate	Av
4996	15	F	7.161276	1.024644	5.312684	10.224924	4	0.427209	Moderate	Exc
4997	14	M	3.444383	2.877972	9.227726	4.059322	4	0.002893	High	
4998	18	F	7.866525	2.395839	4.317831	10.657076	2	0.612063	High	Av
4999	18	M	3.389362	1.375646	8.693171	6.977589	5	0.952662	Moderate	Exc

5000 rows × 10 columns

3. Mengecek dimensi dataset

Langkah berikutnya adalah memeriksa ukuran dataset, yaitu jumlah baris (observasi) dan kolom (variabel). Informasi ini penting untuk mengetahui ruang lingkup data yang akan dianalisis, termasuk mengevaluasi apakah jumlah observasi cukup besar untuk dilakukan analisis statistik.

```
# Mengecek dimensi data
df.shape
```

```
(5000, 10)
```

4. Melihat daftar nama kolom

Untuk memastikan bahwa seluruh kolom dalam dataset telah terbaca dan dinamai dengan benar, dilakukan pengecekan terhadap nama-nama kolom. Hal ini juga bermanfaat dalam memastikan bahwa tidak ada kolom yang salah ketik, kosong, atau duplikat.

```
# Cek kolom
df.columns
```

```
Index(['Age', 'Gender', 'Social_Media_Hours', 'Exercise_Hours', 'Sleep_Hours',
       'Screen_Time_Hours', 'Survey_Stress_Score', 'Wearable_Stress_Score',
       'Support_System', 'Academic_Performance'],
      dtype='object')
```

5. Mengecek tipe data tiap variabel

   Selanjutnya dilakukan pemeriksaan terhadap tipe data masing-masing kolom. Hal ini bertujuan untuk memastikan bahwa variabel-variabel numerik benar-benar dikenali sebagai numerik, dan variabel kategorik dikenali sebagai objek atau faktor. Kesalahan pada tipe data bisa mengganggu proses analisis statistik, misalnya ketika akan dilakukan regresi, visualisasi, atau pemodelan.

```
# Mengecek tipe data setiap variabel
df.dtypes
```

	0
Age	int64
Gender	object
Social_Media_Hours	float64
Exercise_Hours	float64
Sleep_Hours	float64
Screen_Time_Hours	float64
Survey_Stress_Score	int64
Wearable_Stress_Score	float64
Support_System	object
Academic_Performance	object

dtype: object

```
# Mengecek dan mengoreksi tipe variabel
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   5000 non-null   int64
 1   Gender                5000 non-null   object
 2   Social_Media_Hours    5000 non-null   float64
 3   Exercise_Hours        5000 non-null   float64
 4   Sleep_Hours           5000 non-null   float64
 5   Screen_Time_Hours     5000 non-null   float64
 6   Survey_Stress_Score   5000 non-null   int64
 7   Wearable_Stress_Score 5000 non-null   float64
 8   Support_System        5000 non-null   object
 9   Academic_Performance  5000 non-null   object
dtypes: float64(5), int64(2), object(3)
memory usage: 390.8+ KB
```

6. Menampilkan statistik deskriptif dasar

Sebagai penutup dari bagian eksplorasi awal, dilakukan perhitungan statistik deskriptif untuk seluruh variabel numerik. Statistik ini mencakup nilai minimum, maksimum, mean (rata-rata), standar deviasi, serta kuartil. Tujuannya adalah untuk memahami distribusi dan penyebaran data awal, serta untuk mendeteksi kemungkinan outlier atau nilai ekstrim.

```
# Statistik deskriptif
df.describe()
```

	Age	Social_Media_Hours	Exercise_Hours	Sleep_Hours	Screen_Time_Hours	Survey_Stress_Score	Wearable_Stress_Score
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	15.493200	4.932081	1.498151	7.057370	7.068630	3.015800	0.496618
std	1.715151	2.853928	0.873984	1.722211	2.883494	1.414762	0.289768
min	13.000000	0.000528	0.000473	4.001515	2.000481	1.000000	0.000102
25%	14.000000	2.473150	0.734431	5.611836	4.574327	2.000000	0.244615
50%	16.000000	4.898176	1.483432	7.068874	7.118979	3.000000	0.500404
75%	17.000000	7.369195	2.276089	8.519411	9.526335	4.000000	0.749929
max	18.000000	9.995052	2.999774	9.999229	11.999010	5.000000	0.999812

3.1.2 Cek *Misssing Value* dan Duplikasi Data

Sebelum melanjutkan ke tahap analisis yang lebih kompleks, penting untuk memastikan bahwa data yang digunakan berada dalam kondisi bersih dan layak pakai. Dua aspek penting yang harus diperiksa adalah adanya data yang hilang (missing value) dan adanya data yang tercatat lebih dari satu kali (duplikasi). Kedua hal ini dapat mempengaruhi kualitas hasil analisis dan dapat menyebabkan bias atau kesalahan interpretasi jika tidak ditangani dengan benar.

1. Pengecekan missing value

Langkah pertama adalah memeriksa apakah terdapat nilai kosong atau hilang (missing) pada setiap kolom di dalam dataset. Hal ini dilakukan menggunakan fungsi .isnull().sum() dalam Python.

```
# Cek apakah ada data kosong
print("\nMissing values:")
print(df.isnull().sum())
```

```
Missing values:
Age                    0
Gender                 0
Social_Media_Hours     0
Exercise_Hours         0
Sleep_Hours            0
Screen_Time_Hours      0
Survey_Stress_Score    0
Wearable_Stress_Score  0
Support_System         0
Academic_Performance   0
dtype: int64
```

Dari output tersebut, terlihat bahwa setiap kolom memiliki nilai 0 untuk jumlah missing value, yang artinya tidak ada data yang hilang pada seluruh kolom. Ini menunjukkan bahwa dataset yang digunakan berada dalam kondisi lengkap dan tidak memerlukan proses imputasi data atau penghapusan observasi karena kekosongan data.

2. Pengecekan duplikasi data

Setelah memastikan tidak ada nilai kosong, langkah selanjutnya adalah memeriksa apakah terdapat observasi (baris) yang identik atau terduplikasi di dalam dataset. Hal ini dilakukan menggunakan kombinasi dari beberapa fungsi:

- df.duplicated().sum() digunakan untuk menghitung jumlah baris yang sama persis dengan baris sebelumnya.
- df.shape ditampilkan untuk memverifikasi jumlah total observasi.

```
# Mengecek apakah ada duplikasi data?
print(df.shape)
df.duplicated().sum()
```

```
(5000, 10)
np.int64(0)
```

Berdasarkan hasil yang diperoleh, jumlah baris duplikat adalah 0, yang berarti tidak ada observasi yang tercatat lebih dari satu kali dalam dataset. Dengan demikian, setiap baris merupakan data unik yang mewakili satu responden yang berbeda.

3.1.3 Statistik Deskriptif Tiap Variabel

Setelah memastikan bahwa struktur dataset telah terbaca dengan baik, bersih dari nilai kosong dan duplikasi, serta tidak mengandung variabel non-analitis seperti User_ID, langkah selanjutnya adalah melakukan analisis statistik deskriptif terhadap masing-masing variabel. Tujuan dari tahap ini adalah untuk memperoleh gambaran umum mengenai distribusi, rentang, dan kecenderungan nilai dari setiap variabel yang ada dalam dataset.

Analisis ini dilakukan secara terpisah untuk variabel numerik dan kategorik, mengingat perbedaan sifat datanya. Hasil statistik deskriptif memberikan informasi penting untuk mengenali pola dasar dalam data, mendeteksi potensi nilai ekstrim (outlier), serta memahami bentuk distribusi yang nantinya akan mempengaruhi pemilihan metode analisis lanjutan.

1. Statistik deskriptif variabel numerik

   Variabel numerik dalam dataset ini antara lain: Age, Social_Media_Hours, Exercise_Hours, Sleep_Hours, Screen_Time_Hours, Survey_Stress_Score, dan Wearable_Stress_Score.

```
# Statistik deskriptif
df.describe()
```

	Age	Social_Media_Hours	Exercise_Hours	Sleep_Hours	Screen_Time_Hours	Survey_Stress_Score	Wearable_Stress_Score
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	15.493200	4.932081	1.498151	7.057370	7.068630	3.015800	0.496618
std	1.715151	2.853928	0.873984	1.722211	2.883494	1.414762	0.289768
min	13.000000	0.000528	0.000473	4.001515	2.000481	1.000000	0.000102
25%	14.000000	2.473150	0.734431	5.611836	4.574327	2.000000	0.244615
50%	16.000000	4.898176	1.483432	7.068874	7.118979	3.000000	0.500404
75%	17.000000	7.369195	2.276089	8.519411	9.526335	4.000000	0.749929
max	18.000000	9.995052	2.999774	9.999229	11.999010	5.000000	0.999812

2. Statistik deskriptif variabel kategorik

   Variabel kategorik dalam dataset meliputi: Gender, Support_System, dan Academic_Performance.

```
# Mengecek noise di python bisa dengan melihat distribusi dari variabel kategorik
df['Gender'].value_counts()
```

Gender	count
F	2528
M	2472

dtype: int64

```
df['Support_System'].value_counts()
```

Support_System	count
Moderate	1696
High	1677
Low	1627

dtype: int64

```
df['Academic_Performance'].value_counts()
```

Academic_Performance	count
Poor	1293
Good	1271
Average	1234
Excellent	1202

dtype: int64

Analisis dilakukan dengan menghitung frekuensi dan proporsi masing-masing kategori. Hasilnya antara lain:

- Gender: Terdapat dua kategori utama yaitu laki-laki (M) dan perempuan (F). Distribusi cukup seimbang, meskipun terdapat sedikit perbedaan jumlah antar kelompok.
- Support_System: Dukungan sosial terbagi menjadi tiga tingkat: Low, Moderate, dan High. Sebagian besar responden mengaku memiliki dukungan sosial tingkat Moderate, sementara hanya sebagian kecil yang memiliki Low support.
- Academic_Performance: Kategori performa akademik meliputi Poor, Average, Good, dan Excellent. Sebaran menunjukkan bahwa mayoritas responden berada pada tingkat Average dan Good, sedangkan responden dengan performa Poor dan Excellent jumlahnya lebih sedikit.

3.1.4 Cek Nilai Uniq tiap Variabel Kategorik

Sebelum melakukan analisis eksploratif dan visualisasi data, penting untuk memahami komposisi dari setiap variabel kategorik. Salah satu langkah dasar yang dilakukan adalah mengecek jumlah dan jenis nilai unik (unique values) yang terdapat pada masing-masing kolom kategorik. Tujuannya adalah untuk:

- Mengetahui berapa banyak kategori yang dimiliki oleh setiap variabel.
- Memastikan bahwa tidak ada kategori yang tertulis dengan format tidak konsisten (misalnya huruf kapitalisasi yang berbeda).
- Memahami distribusi awal sebagai dasar untuk membuat visualisasi dan analisis perbandingan.

Gender

F	2528
M	2472

dtype: int64

Support_System

Moderate	1696
High	1677
Low	1627

dtype: int64

Academic_Performance

Poor	1293
Good	1271
Average	1234
Excellent	1202

dtype: int64

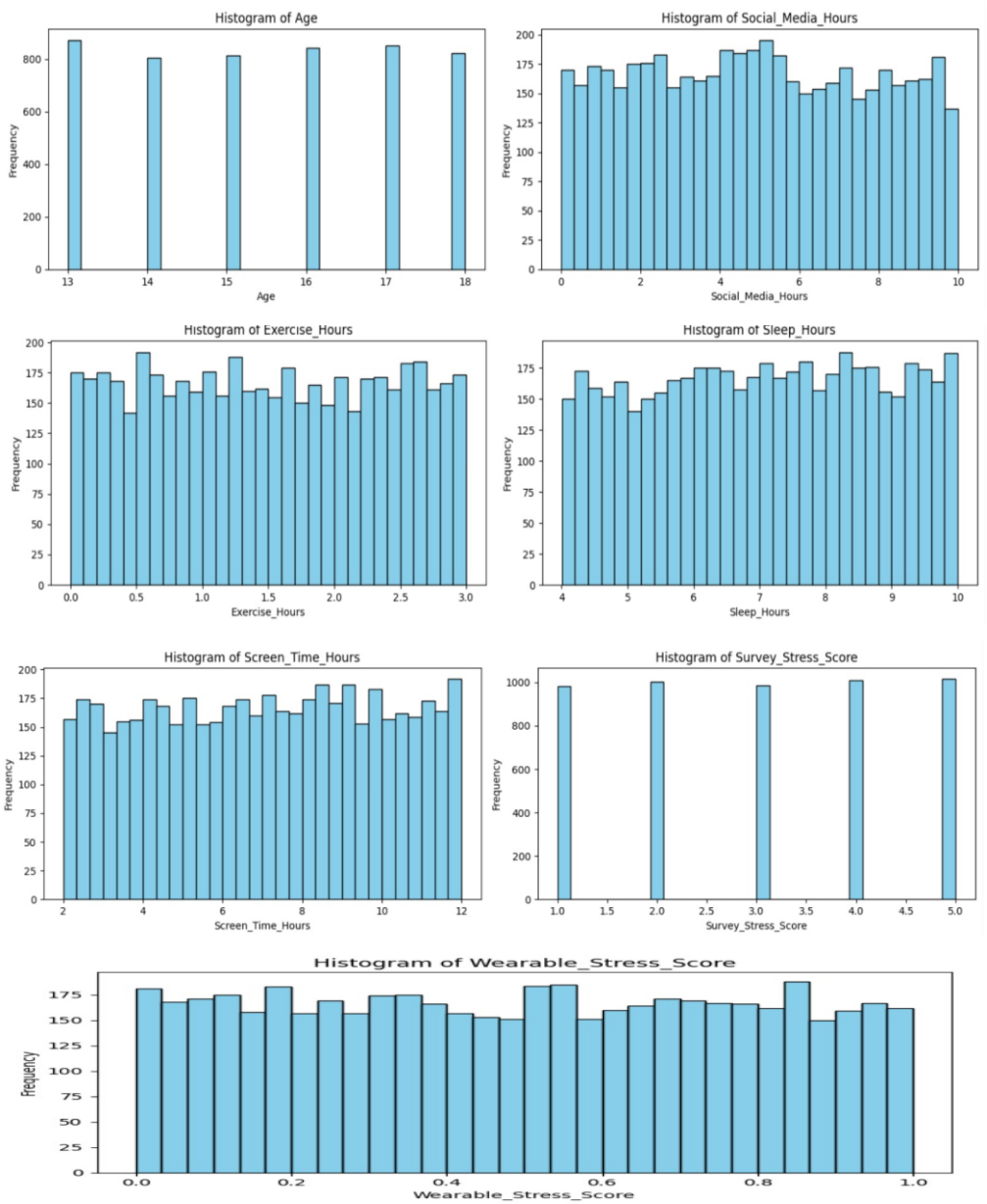
3.1.5 Visualisasi Distribusi Data

Visualisasi data merupakan salah satu langkah dalam proses analisis data. Beberapa bentuk visualisasi yang umum digunakan adalah histogram dan boxplot. Histogram digunakan untuk melihat distribusi frekuensi dari suatu variabel, sedangkan boxplot memberikan informasi tentang penyebaran data dan mendeteksi outlier.

Di bagian ini akan disajikan analisis eksploratif terhadap data numerik menggunakan histogram dan boxplot, dengan fokus pada variabel-variabel seperti durasi olahraga, waktu tidur, screen time, penggunaan media sosial, skor stres, dan usia.

a. Histogram

Dalam analisis ini, dilakukan eksplorasi data terhadap variabel-variabel numerik menggunakan histogram. Tujuan dari histogram adalah untuk memahami distribusi frekuensi dari setiap variabel numerik yang ada pada dataset. Berikut ini penjelasan masing-masing variabel berdasarkan hasil visualisasi histogram:



1. Exercise_Hours

Histogram Exercise_Hours menunjukkan sebaran waktu olahraga peserta. Nilainya berkisar dari 0 hingga 3 jam. Distribusinya cenderung merata, artinya peserta memiliki variasi waktu olahraga yang cukup beragam dan tidak terkonsentrasi di satu rentang waktu tertentu. Ini bisa menandakan bahwa tidak ada pola khusus dalam kebiasaan olahraga para peserta. Tidak tampak adanya outlier. Nilai maksimal adalah 3 jam dan semua data berada dalam rentang wajar (0–3 jam). Bentuk distribusi hampir uniform (merata). Frekuensi setiap interval

cukup seragam, artinya waktu olahraga tidak terkonsentrasi pada satu titik tertentu.

2. Sleep_Hours

   Histogram Sleep_Hours memiliki rentang antara 4 hingga 10 jam. Meskipun ada sedikit fluktuasi, distribusinya tetap relatif merata. Ini mengindikasikan bahwa sebagian besar peserta tidur dalam rentang waktu tersebut, tanpa kecenderungan ekstrem tidur terlalu sedikit atau terlalu lama. Bentuk distribusi hampir uniform, dengan sedikit fluktuasi. Distribusinya tidak terlalu mencolok condong ke kiri atau kanan.

3. Screen_Time_Hours

   Variabel Screen_Time_Hours memperlihatkan durasi penggunaan layar (screen time) antara 2 hingga 12 jam. Pola distribusinya juga relatif merata, dengan sedikit lonjakan di beberapa interval, seperti sekitar 12 jam. Hal ini menunjukkan bahwa banyak peserta yang menghabiskan waktu cukup lama di depan layar setiap harinya. Bentuk distribusi hampir uniform juga, meski ada sedikit kecenderungan naik di interval tinggi (sekitar 12 jam), artinya beberapa peserta punya screen time sangat tinggi.

4. Survey_Stress_Score

   Variabel ini merupakan skor stres berdasarkan survei, dengan nilai diskrit dari 1 sampai 5. Histogramnya menunjukkan distribusi yang hampir seragam, artinya jumlah peserta yang memberikan skor stres dari 1 hingga 5 hampir sama banyak. Ini bisa mengindikasikan bahwa tingkat stres subjektif berdasarkan persepsi cukup tersebar merata di antara peserta. Merata (uniform discrete). Jumlah responden di setiap kategori hampir sama, tidak ada nilai yang dominan.

5. Age

   Histogram usia (Age) berkisar antara 13 hingga 18 tahun. Distribusinya cukup seimbang di setiap usia, menunjukkan bahwa dataset mencakup perwakilan usia remaja yang cukup merata. Ini penting karena bisa memperkuat kesimpulan yang diambil untuk populasi usia tersebut. Jumlah responden di setiap kategori hampir sama, tidak ada nilai yang dominan. Relatif merata. Tidak ada usia yang sangat dominan atau sangat rendah frekuensinya.

6. Social_Media_Hours

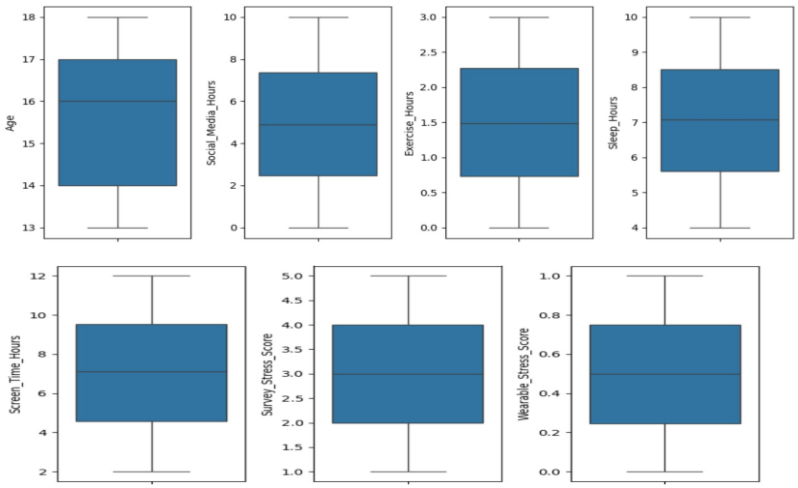
Jam penggunaan media sosial juga cukup bervariasi, dengan rentang antara 0 hingga 10 jam per hari. Tidak ada pola dominan, namun terlihat bahwa cukup banyak peserta yang menggunakan media sosial dalam jumlah yang cukup signifikan. Ini menyoroti pentingnya media sosial dalam keseharian mereka. Bentuk distribusi relatif merata. Tidak ada usia yang sangat dominan atau sangat rendah frekuensinya.

7. Wearable_Stress_Score

Variabel ini mengukur stres berdasarkan data dari perangkat wearable, dengan nilai antara 0 hingga 1. Distribusinya merata di seluruh rentang, menandakan bahwa skor wearable stress cukup bervariasi antar peserta. Perlu diteliti lebih lanjut apakah skor ini memiliki hubungan dengan variabel lain seperti screen time, sleep, atau media sosial. Hampir uniform, artinya perangkat wearable memberikan skor stres yang sangat bervariasi dan tersebar merata.

b. Boxplot

Boxplot memberikan informasi mengenai nilai minimum, kuartil bawah, median, kuartil atas, dan maksimum, serta membantu dalam mendeteksi adanya outlier. Berikut ini penjelasan masing-masing variabel berdasarkan hasil visualisasi Boxplot:



1. Age

Distribusi cukup simetris. Median berada di sekitar usia 16 tahun. Tidak terdapat outlier. Penyebaran usia merata dari 13–18 tahun hal ini memperkuat validitas analisis karena sampel mewakili rentang usia remaja.

2. Social_Media_Hours

   Median penggunaan sekitar 5 jam/hari. Rentang data dari 0 hingga 10 jam. Tidak terlihat outlier. Data cenderung simetris, menandakan variasi jam penggunaan media sosial yang cukup seimbang antar siswa.

3. Exercise_Hours

   Median sekitar 1.5 jam. Tidak ada pencilan, semua data berada dalam rentang 0–3 jam. Rentang interkuartilnya (IQR) agak lebar → artinya tingkat aktivitas fisik antar siswa cukup bervariasi.

4. Sleep_Hours

   Median waktu tidur sekitar 7 jam. Tidak ada outlier. Sebaran cukup seimbang dari 4–10 jam. Ini menandakan kebiasaan tidur siswa cenderung dalam batas normal, walau ada yang tidur hanya 4 jam.

5. Screen_Time_Hours

   Median screen time sekitar 7 jam. Tidak ada outlier mencolok. Rentangnya. dari 2 sampai 12 jam → beberapa siswa punya screen time yang sangat tinggi.

6. Survey_Stress_Score

   Median skor stres sekitar 3 (nilai tengah dari skala 1–5). Distribusinya cukup simetris, tidak ada pencilan. Artinya tingkat stres berdasarkan persepsi cukup berimbang antar siswa.

7. Wearable_Stress_Score

   Median wearable stress score sekitar 0.5. Tidak ada outlier. Penyebaran data dari 0 hingga 1 cukup merata. Bisa diartikan bahwa wearable device mengukur stres dalam cara yang lebih terdistribusi dibanding persepsi pribadi.

3.1.1 Cek Outlier

bertujuan untuk mendeteksi baris data yang mengandung outlier berdasarkan nilai Z-score. Berikut penjabaran langkah dan outputnya:

```
from scipy.stats import zscore

# Ambil hanya kolom numerik
numerical_cols = df.select_dtypes(include=['float64', 'int64'])

# Hitung Z-score
z_scores = zscore(numerical_cols, nan_policy='omit')

# Tandai outlier jika Z-score lebih dari 3 atau kurang dari -3
outliers = (abs(z_scores) > 3)

# Hitung jumlah baris (row) yang memiliki setidaknya satu outlier
jumlah_outlier_baris = outliers.any(axis=1).sum()

print("Jumlah baris dengan minimal satu outlier:", jumlah_outlier_baris)
```

```
Jumlah baris dengan minimal satu outlier: 0
```

Kesimpulan berdasarkan hasil yang di dapatkan dari codingan diatas:

1. Validasi Data Bersih
   - Hasil menunjukkan bahwa dataset bersih dari pencilan ekstrem.
   - Tidak diperlukan proses pembersihan atau transformasi tambahan sebelum melanjutkan ke analisis lanjutan (misal regresi atau klasifikasi).

2. Z-Score sebagai Detektor Outlier
   - Z-score adalah metode statistik yang andal untuk data dengan distribusi mendekati normal.
   - Skor Z menunjukkan seberapa jauh suatu nilai dari rata-rata dalam satuan standar deviasi.
3. Kriteria Z > 3

   Threshold 3 (atau -3) adalah aturan umum dalam statistik: nilai di luar 3 standar deviasi dari rata-rata dianggap sangat jarang dan dianggap outlier.
4. Konsistensi dengan Boxplot
   - Hasil ini mendukung visualisasi sebelumnya (boxplot) yang juga tidak menunjukkan pencilan data.

- Ini memperkuat keyakinan bahwa data bisa digunakan langsung dalam pemodelan statistik.

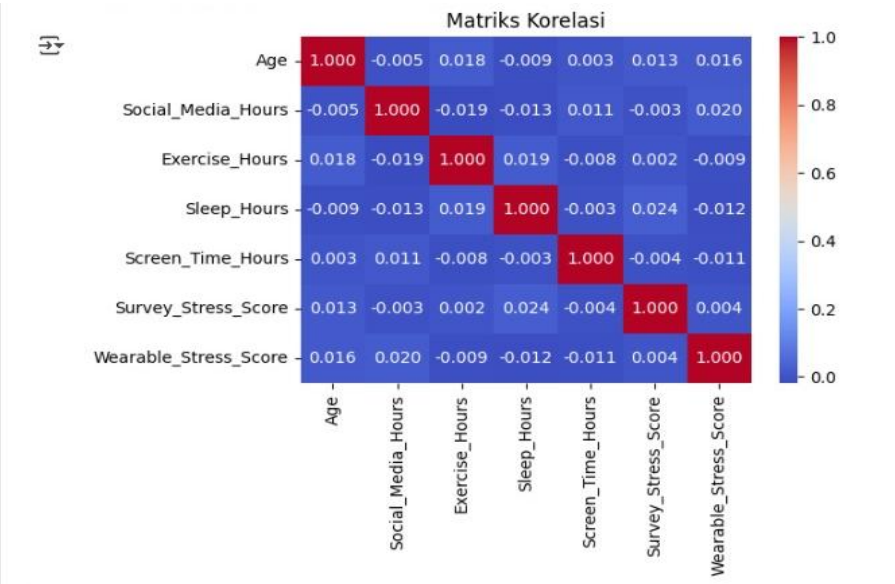
3.1.6 Cek Korelasi antar Variabel Numerik

Analisis korelasi antar variabel numerik menggunakan matriks korelasi berbentuk heatmap. Tujuan dari analisis ini adalah untuk mengidentifikasi apakah terdapat hubungan linier yang kuat antara variabel-variabel yang diamati.

Korelasi dihitung menggunakan metode Pearson, yang mengukur kekuatan dan arah hubungan linier antara dua variabel. Nilai korelasi berkisar antara -1 hingga +1, di mana:

- +1 menunjukkan korelasi positif sempurna
- -1 menunjukkan korelasi negatif sempurna
- 0 menunjukkan tidak ada hubungan linier

Matriks korelasi dapat divisualisasikan menggunakan heatmap yang menunjukkan kekuatan hubungan antar variabel dalam skala warna (biru untuk negatif, merah untuk positif). Berikut adalah visualisasi dan interpretasi dari hasil matriks korelasi:



3.1.7 Analisis Hubungan antar Variabel

Pasangan Variabel	Nilai r	Interpretasi Singkat
Age vs lainnya	±0.01 – 0.03	Hampir tak berkorelasi; usia remaja di sampel tidak memengaruhi kebiasaan harian.
Social_Media_Hours vs Screen_Time_Hours	0.033	Positif sangat lemah – wajar karena screen-time sebagian besar berasal dari media sosial, tapi efeknya kecil sekali.
Exercise_Hours vs Sleep_Hours	0.019	Hubungan positif lemah; makin rajin olahraga sedikit cenderung tidur lebih lama, namun tidak signifikan.
Sleep_Hours vs Survey_Stress_Score	-0.024	Negatif sangat lemah; tidur lebih lama sedikit terkait stres survei lebih rendah, tapi efek nyaris nol.
Survey_Stress_Score vs Wearable_Stress_Score	0.004	Kedua metode pengukuran stres hampir independen.
Korelasi terkuat di seluruh matriks	≈ 0.04	Seluruh nilai < 0.05 ⇒ tidak ada korelasi praktis.

Maka dapat disimpulkan, bahwa:

- Nilai r mendekati 0 menandakan variabel-variabel cenderung independen secara linier.
- Rendahnya korelasi juga bisa berarti responden homogen atau variabel belum cukup variatif.
- Multikolinearitas nyaris tidak ada → baik untuk regresi multipel karena prediktor tidak saling tumpang-tindih.

Jadi Heatmap korelasi mengonfirmasi bahwa di dalam dataset ini tidak terdapat hubungan linier kuat antara variabel numerik apa pun (< |0.05|). Hal ini memvalidasi bahwa analisis lanjutan tidak perlu khawatir tentang multikolinearitas, tetapi juga menegaskan perlunya pendekatan selain korelasi

linier untuk mengeksplorasi hubungan mendalam—misalnya model non-linier, teknik machine-learning, atau analisis kelompok spesifik.

3.2 Strategi Model Building

Membangun Full Model:

Model regresi linear berganda ini dibuat untuk memprediksi tingkat stres berdasarkan data yang diperoleh dari perangkat wearable (Wearable_Stress_Score). Beberapa variabel yang digunakan mencakup usia (Age), jenis kelamin (Gender), penggunaan media sosial (Social_Media_Hours), aktivitas olahraga (Exercise_Hours), durasi tidur (Sleep_Hours), waktu layar (Screen_Time_Hours), hasil survei stres (Survey_Stress_Score), dukungan sosial (Support_System), serta performa akademik (Academic_Performance). Model ini juga memperhitungkan hubungan non-linier seperti kuadrat dari variabel-variabel tertentu, serta interaksi antar variabel, contohnya pengaruh olahraga yang berbeda tergantung pada tingkat dukungan sosial. Dengan pendekatan ini, model dapat memahami hubungan yang lebih kompleks di antara faktor-faktor tersebut. Meski demikian, evaluasi lanjutan diperlukan agar model tidak terlalu sempurna dalam data tertentu dan hasilnya bisa diterapkan secara luas.

```
# Model full
model_full <- lm(Wearable_Stress_Score ~
                   Age + Gender + Social_Media_Hours + Exercise_Hours + Sleep_Hours + Screen_Time_Hours +
                   Survey_Stress_Score + Support_System + Academic_Performance + I(Social_Media_Hours^2) +
                   I(Exercise_Hours^2) + I(Sleep_Hours^2) + I(Screen_Time_Hours^2) +
                   I(Survey_Stress_Score^2) + Age:Gender + Age:Social_Media_Hours + Age:Exercise_Hours +
                   Age:Sleep_Hours + Age:Screen_Time_Hours + Age:Survey_Stress_Score +
                   Age:Support_System + Age:Academic_Performance + Gender:Social_Media_Hours +
                   Gender:Exercise_Hours + Gender:Sleep_Hours + Gender:Screen_Time_Hours +
                   Gender:Survey_Stress_Score + Gender:Support_System + Gender:Academic_Performance +
                   Social_Media_Hours:Exercise_Hours + Social_Media_Hours:Sleep_Hours +
                   Social_Media_Hours:Screen_Time_Hours + Social_Media_Hours:Survey_Stress_Score +
                   Social_Media_Hours:Support_System + Social_Media_Hours:Academic_Performance +
                   Exercise_Hours:Sleep_Hours + Exercise_Hours:Screen_Time_Hours +
                   Exercise_Hours:Survey_Stress_Score + Exercise_Hours:Support_System +
                   Exercise_Hours:Academic_Performance + Sleep_Hours:Screen_Time_Hours +
                   Sleep_Hours:Survey_Stress_Score + Sleep_Hours:Support_System +
                   Sleep_Hours:Academic_Performance + Screen_Time_Hours:Survey_Stress_Score +
                   Screen_Time_Hours:Support_System + Screen_Time_Hours:Academic_Performance +
                   Survey_Stress_Score:Support_System + Survey_Stress_Score:Academic_Performance +
                   Support_System:Academic_Performance, data = data)
```

Secara keseluruhan, model ini sangat lengkap dan dibuat untuk menilai dampak yang saling terkait dari berbagai aspek kehidupan mahasiswa terhadap stres. Meski demikian, karena tingginya kompleksitas dengan banyak variabel dan interaksi antarvariabel, diperlukan evaluasi lebih lanjut seperti pemilihan model, uji multikolinearitas, serta validasi model agar memastikan model tidak terlalu sempurna dalam data tertentu dan hasilnya bisa diterapkan secara umum.

Seleksi Model:

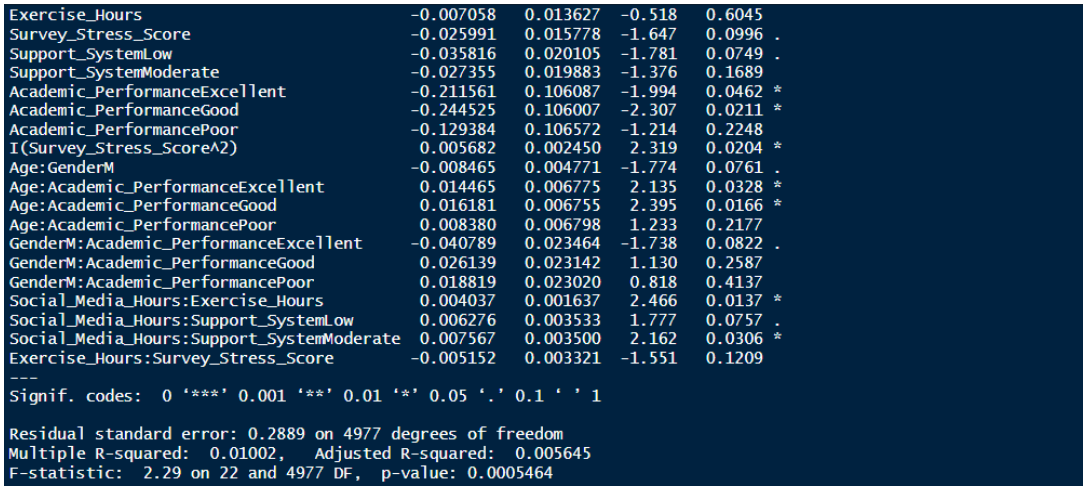
Model regresi awal yang dibuat adalah model penuh dengan melibatkan semua variabel prediktor, termasuk efek kuadrat dan interaksi dua arah. Meskipun model ini secara keseluruhan menunjukkan signifikansi berdasarkan uji F, terdapat banyak variabel yang tidak signifikan berdasarkan uji t, yang menunjukkan kemungkinan adanya kompleksitas berlebihan atau multikolinearitas antar prediktor. Karena itu, dilakukan proses pemilihan model untuk menyederhanakan model serta meningkatkan akurasi prediksi dan kemudahan dalam memahami hasilnya.

Pemilihan model dilakukan dengan dua cara. Pertama, menggunakan metode stepwise selection berdasarkan kriteria AIC (Akaike Information Criterion) yang secara otomatis memilih kombinasi variabel terbaik dengan mempertimbangkan keseimbangan antara tingkat kompleksitas model dan kualitas penjelasan variabel terikat. Kedua, dilakukan pendekatan manual dengan menganalisis multikolinearitas antar variabel menggunakan nilai Variance Inflation Factor (VIF). Variabel dengan nilai VIF tinggi (biasanya lebih dari 5 atau 10) menunjukkan adanya korelasi yang kuat dengan variabel lain, sehingga perlu dipertimbangkan untuk dihilangkan atau direvisi agar menghindari distorsi dalam estimasi koefisien model akhir.

1. Metode Stepwise

   Untuk menyederhanakan model dan meningkatkan interpretabilitas, dilakukan seleksi variabel menggunakan metode **Stepwise Regression** berbasis **Akaike Information Criterion (AIC)**. Proses dimulai dari model penuh (full model) yang mencakup variabel prediktor, termasuk efek linear kuadratik dan interaksi antar variabel. Hasilnya adalah model yang lebih efisien, hanya mempertahankan variabel-variabel signifikan yang benar-benar berkontribusi terhadap skor stres pada perangkat wearable, serta mengurangi risiko overfitting akibat kompleksitas model awal.

```
> summary(step_model_aic)

Call:
lm(formula = Wearable_Stress_Score ~ Age + Gender + Social_Media_Hours +
    Exercise_Hours + Survey_Stress_Score + Support_System + Academic_Performance +
    I(Survey_Stress_Score^2) + Age:Gender + Age:Academic_Performance +
    Gender:Academic_Performance + Social_Media_Hours:Exercise_Hours +
    Social_Media_Hours:Support_System + Exercise_Hours:Survey_Stress_Score,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54327 -0.24900  0.00256  0.24940  0.56308

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       0.593612   0.090448   6.563 5.81e-11 ***
Age                              -0.002471   0.005376  -0.460   0.6459
GenderM                           0.123018   0.075787   1.623   0.1046
Social_Media_Hours               -0.008605   0.003542  -2.430   0.0152 *
```



```
Exercise_Hours                          -0.007058   0.013627  -0.518   0.6045
Survey_Stress_Score                     -0.025991   0.015778  -1.647   0.0996 .
Support_SystemLow                       -0.035816   0.020105  -1.781   0.0749 .
Support_SystemModerate                  -0.027355   0.019883  -1.376   0.1689
Academic_PerformanceExcellent           -0.211561   0.106087  -1.994   0.0462 *
Academic_PerformanceGood                -0.244525   0.106007  -2.307   0.0211 *
Academic_PerformancePoor                -0.129384   0.106572  -1.214   0.2248
I(Survey_Stress_Score^2)                 0.005682   0.002450   2.319   0.0204 *
Age:GenderM                             -0.008465   0.004771  -1.774   0.0761 .
Age:Academic_PerformanceExcellent        0.014465   0.006775   2.135   0.0328 *
Age:Academic_PerformanceGood             0.016181   0.006755   2.395   0.0166 *
Age:Academic_PerformancePoor             0.008380   0.006798   1.233   0.2177
GenderM:Academic_PerformanceExcellent   -0.040789   0.023464  -1.738   0.0822 .
GenderM:Academic_PerformanceGood         0.026139   0.023142   1.130   0.2587
GenderM:Academic_PerformancePoor         0.018819   0.023020   0.818   0.4137
Social_Media_Hours:Exercise_Hours        0.004037   0.001637   2.466   0.0137 *
Social_Media_Hours:Support_SystemLow     0.006276   0.003533   1.777   0.0757 .
Social_Media_Hours:Support_SystemModerate 0.007567  0.003500   2.162   0.0306 *
Exercise_Hours:Survey_Stress_Score      -0.005152   0.003321  -1.551   0.1209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2889 on 4977 degrees of freedom
Multiple R-squared:  0.01002,   Adjusted R-squared:  0.005645
F-statistic:  2.29 on 22 and 4977 DF,  p-value: 0.0005464
```

Berdasarkan hasil analisis regresi linear dengan metode stepwise AIC, model akhir untuk memprediksi Wearable_Stress_Score mengandung 22 variabel prediktor, termasuk beberapa interaksi dan efek kuadrat. Nilai AIC sebesar -12392.09 menunjukkan bahwa model tersebut cukup baik dalam menyeimbangkan kompleksitas dan kesesuaian dengan data. Namun demikian, nilai R-squared hanya 0.01002 dan Adjusted R-squared sebesar 0.005645, yang menunjukkan bahwa model hanya mampu menjelaskan sekitar 1% variasi dalam Wearable Stress Score. Dengan demikian, daya prediksi model secara keseluruhan tergolong sangat rendah.

Beberapa variabel yang berpengaruh signifikan ($p < 0.05$) dalam model adalah sebagai berikut:

- `Social_Media_Hours` ($p = 0.0152$), yang memiliki pengaruh negatif signifikan terhadap Wearable Stress Score. Artinya, semakin lama seseorang menggunakan media sosial, semakin rendah nilai stres yang terdeteksi oleh wearable. Ini bisa menunjukkan media sosial sebagai mekanisme untuk mengurangi stres atau mungkin disebabkan oleh ketidakakuratan data.
- `Academic_PerformanceGood` dan `Academic_PerformanceExcellent` memiliki hubungan negatif signifikan terhadap skor stres, yang menunjukkan bahwa siswa dengan kemampuan akademik baik cenderung memiliki tingkat stres lebih rendah sesuai hasil pengukuran wearable.
- `I(Survey_Stress_Score^2)` ($p = 0.0204$) menunjukkan adanya hubungan nonlinier antara skor stres yang dilaporkan secara subjektif dan skor stres yang terdeteksi oleh wearable, mengindikasikan bahwa pengaruh ini tidak bersifat linear.

- Interaksi antara `Social_Media_Hours` dan `Exercise_Hours` (p = 0.0137) menunjukkan pengaruh positif yang signifikan, artinya penggunaan media sosial dalam memengaruhi stres wearable berubah tergantung pada durasi olahraga yang dilakukan.
- Interaksi `Age:Academic_Performance` untuk kategori Good dan Excellent juga menunjukkan pengaruh yang signifikan, mengindikasikan bahwa usia berpengaruh terhadap stres wearable, tetapi tingkat pengaruh ini bisa berubah tergantung pada performa akademik individu.

Beberapa variabel seperti `Gender`, `Exercise_Hours`, `Support_System`, dan interaksi lainnya tidak signifikan secara statistik, namun tetap dimasukkan dalam model karena proses seleksi stepwise AIC menelusuri kontribusi variabel tersebut terhadap informasi keseluruhan model.

Secara keseluruhan, meskipun beberapa variabel dan interaksi menunjukkan pengaruh statistik yang signifikan, hasil R-squared yang rendah mengindikasikan bahwa model ini belum mampu menjelaskan secara baik variasi dalam data stres wearable.

2. Uji Multikolinearitas

Untuk mengenali adanya kemungkinan multikolinearitas antar variabel independen dalam model regresi, dilakukan analisis Variance Inflation Factor (VIF). Nilai VIF digunakan untuk mengetahui sejauh mana suatu variabel berkorelasi tinggi dengan variabel lainnya dalam model. Hasil perhitungan nilai VIF untuk masing-masing variabel tersaji pada Gambar berikut:

```
> # Lihat VIF
> vif(model_full)
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
                                        GVIF Df GVIF^(1/(2*Df))
Age                                4.076952e+01  1        6.385102
Gender                             1.226527e+02  1       11.074866
Social_Media_Hours                 1.334589e+02  1       11.552442
Exercise_Hours                     1.354492e+02  1       11.638265
Sleep_Hours                        1.892139e+02  1       13.755506
Screen_Time_Hours                  1.458669e+02  1       12.077539
Survey_Stress_Score                1.462827e+02  1       12.094737
Support_System                     1.475057e+04  2       11.020524
Academic_Performance               1.722181e+06  3       10.948294
I(Social_Media_Hours^2)            1.591220e+01  1        3.989010
I(Exercise_Hours^2)                1.637079e+01  1        4.046083
I(Sleep_Hours^2)                   8.261489e+01  1        9.089274
I(Screen_Time_Hours^2)             3.082368e+01  1        5.551908
I(Survey_Stress_Score^2)           2.731732e+01  1        5.226598
Age:Gender                         8.533159e+01  1        9.237510
Age:Social_Media_Hours             8.660058e+01  1        9.305943
Age:Exercise_Hours                 8.745180e+01  1        9.351567
Age:Sleep_Hours                    9.935326e+01  1        9.967611
Age:Screen_Time_Hours              9.076515e+01  1        9.527074
Age:Survey_Stress_Score            8.918677e+01  1        9.443874
Age:Support_System                 7.209040e+03  2        9.214449
Age:Academic_Performance           6.070878e+05  3        9.201852
Gender:Social_Media_Hours          5.084686e+00  1        2.254925
```

Gender:Exercise_Hours	5.003534e+00	1	2.236858
Gender:Sleep_Hours	1.894503e+01	1	4.352589
Gender:Screen_Time_Hours	8.070120e+00	1	2.840796
Gender:Survey_Stress_Score	6.647620e+00	1	2.578298
Gender:Support_System	8.280340e+00	2	1.696337
Gender:Academic_Performance	2.022729e+01	3	1.650655
Social_Media_Hours:Exercise_Hours	6.959184e+00	1	2.638026
Social_Media_Hours:Sleep_Hours	2.085362e+01	1	4.566576
Social_Media_Hours:Screen_Time_Hours	1.053964e+01	1	3.246481
Social_Media_Hours:Survey_Stress_Score	8.765425e+00	1	2.960646
Social_Media_Hours:Support_System	2.512786e+01	2	2.238921
Social_Media_Hours:Academic_Performance	1.176724e+02	3	2.213668
Exercise_Hours:Sleep_Hours	2.114296e+01	1	4.598148
Exercise_Hours:Screen_Time_Hours	1.005286e+01	1	3.170624
Exercise_Hours:Survey_Stress_Score	8.656451e+00	1	2.942185
Exercise_Hours:Support_System	2.407362e+01	2	2.215059
Exercise_Hours:Academic_Performance	1.141638e+02	3	2.202528
Sleep_Hours:Screen_Time_Hours	2.417246e+01	1	4.916549
Sleep_Hours:Survey_Stress_Score	2.275839e+01	1	4.770575
Sleep_Hours:Support_System	3.652276e+02	2	4.371605
Sleep_Hours:Academic_Performance	6.934693e+03	3	4.366880
Screen_Time_Hours:Survey_Stress_Score	1.157820e+01	1	3.402676
Screen_Time_Hours:Support_System	6.498596e+01	2	2.839258
Screen_Time_Hours:Academic_Performance	5.223932e+02	3	2.837916
Survey_Stress_Score:Support_System	4.297466e+01	2	2.560372
Survey_Stress_Score:Academic_Performance	2.756779e+02	3	2.551136
Support_System:Academic_Performance	2.400203e+02	6	1.578898

Hasil analisis VIF menunjukkan bahwa mayoritas variabel dalam model memiliki nilai GVIF^(1/(2*Df)) dibawah batas umum 10, yang menunjukkan bahwa tidak ada masalah multikolinearitas yang serius antar variabel prediktor. Namun, beberapa variabel seperti Sleep_Hours (13.755506), Survey_Stress_Score (12.094737), dan Social_Media_Hours (11.552442) memiliki nilai yang cukup tinggi, sehingga perlu diwaspadai sebagai kemungkinan penyebab multikolinearitas. Nilai GVIF yang tinggi dapat muncul karena adanya interaksi dan efek kuadrat dalam model, yang secara alami meningkatkan korelasi antar variabel prediktor. Karena itu, meskipun multikolinearitas tidak secara signifikan mengganggu, sebaiknya dipertimbangkan untuk menyederhanakan model atau meninjau ulang struktur interaksi guna meningkatkan kestabilan estimasi koefisien regresi.

Setelah membangun model regresi linear lengkap yang mencakup berbagai variabel utama, interaksi, serta efek non-linear, dilakukan proses penyederhanaan model dengan metode stepwise regression berdasarkan kriteria Akaike Information Criterion (AIC). Proses ini menghasilkan model yang lebih sederhana (model_reduced) yang hanya mempertahankan variabel-variabel prediktor dan interaksi yang memberikan pengaruh signifikan terhadap skor stres yang dideteksi oleh perangkat wearable. Formula dari model yang telah dipilih diperoleh melalui formula(step_model_aic), lalu digunakan kembali untuk menyusun ulang model agar lebih efisien.

```
> formula(step_model_aic)
Wearable_Stress_Score ~ Age + Gender + Social_Media_Hours + Exercise_Hours +
    Survey_Stress_Score + Support_System + Academic_Performance +
    I(Survey_Stress_Score^2) + Age:Gender + Age:Academic_Performance +
    Gender:Academic_Performance + Social_Media_Hours:Exercise_Hours +
    Social_Media_Hours:Support_System + Exercise_Hours:Survey_Stress_Score
> # Misalnya Anda sudah menjalankan stepwise tapi belum menyimpannya:
> model_reduced <- update(model_full, . ~ Age + Gender + Social_Media_Hours +
+                     Exercise_Hours + Survey_Stress_Score + Support_System + Academic_Performance +
+                     I(Survey_Stress_Score^2) + Age:Gender + Age:Academic_Performance +
+                     Gender:Academic_Performance + Social_Media_Hours:Exercise_Hours +
+                     Social_Media_Hours:Support_System + Exercise_Hours:Survey_Stress_Score,
+                   data = data)
```

Selanjutnya adalah cek kembali Variance Inflation Factor (VIF). Berikut adalah hasil perhitungan VIF dari model_reduced:

```
> vif(model_reduced)
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
                                        GVIF Df GVIF^(1/(2*Df))
Age                                5.090832e+00  1        2.256287
Gender                             8.598106e+01  1        9.272597
Social_Media_Hours                 6.117870e+00  1        2.473433
Exercise_Hours                     8.492971e+00  1        2.914270
Survey_Stress_Score                2.983217e+01  1        5.461883
Support_System                     1.604155e+01  2        2.001297
Academic_Performance               5.829373e+05  3        9.139806
I(Survey_Stress_Score^2)           2.701700e+01  1        5.197788
Age:Gender                         8.404919e+01  1        9.167835
Age:Academic_Performance           5.878452e+05  3        9.152586
Gender:Academic_Performance        1.962715e+01  3        1.642390
Social_Media_Hours:Exercise_Hours  6.825861e+00  1        2.612635
Social_Media_Hours:Support_System  2.438238e+01  2        2.222128
Exercise_Hours:Survey_Stress_Score 8.561523e+00  1        2.926008
```

Berdasarkan hasil analisis Variance Inflation Factor (VIF) pada model `model_reduced`, terdapat beberapa variabel yang menunjukkan indikasi multikolinearitas tinggi. Secara umum, jika nilai VIF melebihi 10 (atau GVIF^(1/(2*Df)) > 3. 16), maka dianggap terjadi masalah multikolinearitas yang serius. Dalam hasil ini, beberapa variabel seperti `Gender`, `Survey_Stress_Score`, `I(Survey_Stress_Score^2)`, serta interaksi `Age:Gender` dan `Age:Academic_Performance` memiliki nilai GVIF^(1/(2*Df)) yang mendekati atau bahkan melebihi batas ini, yang menunjukkan bahwa variabel-variabel tersebut saling berkorelasi secara signifikan.

Variabel yang paling mencolok adalah `Academic_Performance` dan interaksi `Age:Academic_Performance` yang memiliki nilai GVIF sangat tinggi (> 500.000) dan nilai GVIF^(1/(2*Df)) di atas 9, menunjukkan adanya multikolinearitas yang sangat kuat dan perlu diwaspadai karena dapat mengganggu stabilitas koefisien regresi. Hal ini berarti, estimasi pengaruh variabel tersebut dalam model bisa menjadi tidak akurat atau sangat rentan terhadap perubahan data.

Namun, beberapa variabel lain seperti `Support_System`, `Social_Media_Hours`, dan interaksi seperti

`Social_Media_Hours:Support_System` masih memiliki nilai GVIF^(1/(2*Df)) di bawah ambang batas (sekitar 2), yang menunjukkan bahwa kontribusi multikolinearitas dari mereka relatif rendah dan tidak terlalu memengaruhi model.

Sebagai lanjutan, dibangun model baru bernama step_model2 sebagai hasil pemodelan lanjutan dari model_full, dengan tujuan mengurangi multikolinearitas yang terdeteksi sebelumnya.

```
> step_model2 <- update(model_full, . ~ Age + Social_Media_Hours +
+                       Exercise_Hours + Support_System +
+                       Gender:Academic_Performance + Social_Media_Hours:Exercise_Hours +
+                       Social_Media_Hours:Support_System + Exercise_Hours:Survey_Stress_Score, data = data)
>
```

```
> vif(step_model2, type = "predictor")
GVIFs computed for predictors
                          GVIF Df GVIF^(1/(2*Df))
Age                  1.0026853  1        1.0013418
Social_Media_Hours   2.9188308  7        1.0795163
Exercise_Hours       6.0533260  4        1.2524179
Support_System       2.7005504  5        1.1044479
Gender               0.6748011  8        0.9757161
Academic_Performance 0.6748011  8        0.9757161
Survey_Stress_Score 10.4931451  2        1.7998090
```

Berdasarkan hasil perhitungan VIF, semua nilai GVIF^(1/(2*Df)) berada di bawah ambang batas multikolinearitas yang biasa digunakan, yaitu 5, bahkan sebagian besar berada di sekitar 1. Hal ini menunjukkan bahwa tidak ada masalah multikolinearitas yang signifikan pada model `step_model2`.

Nilai tertinggi terdapat pada Survey_Stress_Score dengan GVIF sebesar 10.49 dan GVIF^(1/(2*Df)) sebesar 1.799, yang masih dalam batas wajar karena GVIF^(1/(2*Df)) kurang dari 2. Artinya, meskipun Survey_Stress_Score berinteraksi dengan variabel itu sendiri (misalnya melalui kuadrat atau interaksi lain), hubungan antar variabel tidak menyebabkan redundansi yang tinggi dalam model.

Selain itu, variabel Exercise_Hours memiliki GVIF yang cukup tinggi (6.05) namun GVIF^(1/(2*Df)) masih dalam batas aman (1.25), meskipun ia terlibat dalam interaksi dengan dua variabel lain yaitu Social_Media_Hours dan Survey_Stress_Score. Demikian pula, Social_Media_Hours dan Support_System juga terlibat dalam beberapa interaksi, namun nilai GVIF^(1/(2*Df)) masing-masing sebesar 1.08 dan 1.10 yang menunjukkan tidak ada ancaman serius terhadap kestabilan model.

Yang menarik adalah nilai GVIF pada Gender dan Academic_Performance sangat rendah (0.67), bahkan di bawah 1, yang berarti kedua variabel tersebut memiliki hubungan yang sangat lemah atau bahkan negatif terhadap variabel lainnya di dalam model.

Secara keseluruhan, model `step_model2` menunjukkan kestabilan yang baik terhadap multikolinearitas, sehingga pemodelan bisa dilanjutkan tanpa perlu mengeluarkan variabel manapun karena adanya korelasi tinggi antar prediktor.

Model baru yang diperoleh setelah proses penyederhanaan variabel didasarkan pada hasil evaluasi multikolinearitas, signifikansi statistik, dan nilai AIC selama pemilihan variabel secara stepwise, sehingga memastikan model tersebut memiliki keseimbangan antara sifat sederhana dan kemampuan prediksi yang baik.

```
> model <- lm(Wearable_Stress_Score ~ Age + Social_Media_Hours +
+               Exercise_Hours + Support_System +
+               Gender:Academic_Performance + Social_Media_Hours:Exercise_Hours +
+               Social_Media_Hours:Support_System + Exercise_Hours:Survey_Stress_Score,
+             data = data)
```

Berikut adalah hasil analisis regresi linear berganda untuk mengetahui pengaruh masing-masing variabel terhadap tingkat stres yang terdeteksi. Hasil estimasi koefisien, nilai standar error, nilai t, dan signifikansi (p-value) dari masing-masing variabel dalam model ditampilkan pada Gambar berikut:

```
Coefficients: (1 not defined because of singularities)
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             0.503911   0.043164  11.674  < 2e-16
Age                                     0.003079   0.002389   1.289  0.19744
Social_Media_Hours                     -0.008765   0.003545  -2.473  0.01345
Exercise_Hours                         -0.019775   0.010502  -1.883  0.05977
Support_SystemLow                      -0.035644   0.020125  -1.771  0.07660
Support_SystemModerate                 -0.027123   0.019891  -1.364  0.17277
GenderF:Academic_PerformanceAverage    -0.011773   0.016188  -0.727  0.46711
GenderM:Academic_PerformanceAverage    -0.019288   0.016175  -1.192  0.23314
GenderF:Academic_PerformanceExcellent   0.000694   0.016216   0.043  0.96587
GenderM:Academic_PerformanceExcellent  -0.046803   0.016379  -2.858  0.00429
GenderF:Academic_PerformanceGood       -0.004686   0.015832  -0.296  0.76723
GenderM:Academic_PerformanceGood        0.012831   0.016310   0.787  0.43149
GenderF:Academic_PerformancePoor       -0.010655   0.016105  -0.662  0.50825
GenderM:Academic_PerformancePoor              NA         NA      NA       NA
Social_Media_Hours:Exercise_Hours       0.004139   0.001638   2.526  0.01155
Social_Media_Hours:Support_SystemLow    0.006228   0.003536   1.761  0.07829
Social_Media_Hours:Support_SystemModerate 0.007525 0.003502   2.149  0.03170
Exercise_Hours:Survey_Stress_Score     -0.001046   0.001672  -0.626  0.53163
```

Berdasarkan hasil regresi linear berganda pada Gambar, diketahui bahwa beberapa variabel kategorik, seperti Support_System, Gender, dan Academic_Performance, telah dikodekan dalam bentuk variabel dummy.

Dalam metode ini, satu kategori dari setiap variabel kategorik tidak ditampilkan dalam model karena dijadikan kategori referensi (baseline). Penetapan baseline penting untuk menghindari multikolinearitas dan agar koefisien yang ditampilkan dapat diinterpretasikan sebagai perbandingan terhadap baseline tersebut.

Pada output regresi, terlihat bahwa kombinasi GenderM:Academic_PerformancePoor tidak ditampilkan (diberi keterangan "NA"), yang menunjukkan bahwa kombinasi tersebut digunakan sebagai baseline. Oleh karena itu, seluruh kombinasi Gender dan Academic_Performance lainnya diukur relatif terhadap pria dengan performa akademik Poor.

Demikian pula, untuk variabel Support_System, hanya kategori Low dan Moderate yang muncul dalam model, yang berarti bahwa kategori High digunakan sebagai baseline. Maka, koefisien seperti Support_SystemLow = -0.035644 mengindikasikan bahwa skor stres wearable pada peserta dengan dukungan sosial rendah lebih rendah 0.035644 poin dibandingkan peserta dengan dukungan sosial tinggi, dengan asumsi variabel lain tetap.

Dengan demikian, pengetahuan tentang kategori baseline sangat penting untuk menafsirkan arah dan makna dari setiap koefisien dummy yang ada dalam model.

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2893 on 4983 degrees of freedom
Multiple R-squared:  0.006534,  Adjusted R-squared:  0.003344
F-statistic: 2.048 on 16 and 4983 DF,  p-value: 0.008064
```

Berdasarkan output model regresi, model secara keseluruhan signifikan secara statistik (nilai *p* F-test = 0.008064), yang menunjukkan bahwa setidaknya ada satu variabel prediktor yang berkontribusi terhadap model. Namun, kemampuan prediksi model masih sangat rendah, ditunjukkan oleh nilai R-squared = 0.0065 dan Adjusted R-squared = 0.0033, yang berarti model hanya mampu menjelaskan sekitar 0.65% variasi dari skor stres wearable. Jadi, meskipun signifikan, model ini belum cukup kuat dalam menjelaskan hubungan antara variabel-variabel independen dan stres wearable.

Validasi Model:

```
> #Validasi
> train_control <- trainControl(method = "cv", number = 10)
> model_fix <- train(Wearable_Stress_Score ~ Age + Social_Media_Hours +
+                      + Exercise_Hours + Support_System +
+                      + Gender:Academic_Performance + Social_Media_Hours:Exercise_Hou
rs +
+                      + Social_Media_Hours:Support_System + Exercise_Hours:Survey_Str
ess_Score, data = data,
+                    method = "lm", trControl = train_control)
```

Untuk memastikan bahwa model regresi yang dibangun memiliki kemampuan generalisasi yang baik dan tidak mengalami overfitting, dilakukan proses validasi

silang (cross-validation) dengan metode 10-fold cross-validation. Pada tahap ini, data dibagi secara acak menjadi sepuluh subset (fold), di mana model dilatih pada sembilan subset dan diuji pada satu subset secara bergantian hingga seluruh data digunakan sebagai data uji. Model akhir (`model_fix`) dibangun dengan menggunakan beberapa variabel prediktor yang telah diseleksi berdasarkan signifikansi statistik sebelumnya, yaitu `Age`, `Social_Media_Hours`, `Exercise_Hours`, `Support_System`, serta beberapa interaksi dua arah seperti `Gender:Academic_Performance`, `Social_Media_Hours:Exercise_Hours`, `Social_Media_Hours:Support_System`, dan `Exercise_Hours:Survey_Stress_Score`. Prosedur ini bertujuan untuk mengevaluasi stabilitas dan kinerja model secara lebih akurat pada data yang belum pernah dilihat sebelumnya.

```
> model_fix
Linear Regression

5000 samples
   7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4500, 4500, 4500, 4500, 4500, 4500, ...
Resampling results:

  RMSE      Rsquared     MAE
  0.289714  0.004340768  0.250672

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Berdasarkan hasil validasi silang (cross-validation 10-fold) terhadap model regresi linear yang melibatkan 7 prediktor dan 5000 sampel, diperoleh nilai Root Mean Square Error (RMSE) sebesar 0.289714, Mean Absolute Error (MAE) sebesar 0.250672, dan koefisien determinasi (R^2) hanya sebesar 0.0043. Nilai R^2 yang sangat rendah menunjukkan bahwa model hanya mampu menjelaskan sekitar 0,43% variabilitas dari data target (*Wearable_Stress_Score*), yang berarti model memiliki daya prediksi yang sangat lemah. Meskipun kesalahan prediksi relatif kecil berdasarkan nilai RMSE dan MAE, kontribusi model dalam menjelaskan hubungan antar variabel sangat terbatas. Hal ini mengindikasikan bahwa kemungkinan terdapat

faktor-faktor penting lain di luar model yang belum dimasukkan, atau hubungan antar variabel tidak bersifat linear.

3.3 Interpretasi Hasil Regresi Linear

3.3.1. Bentuk Model Regresi

Model regresi linear dibangun menggunakan pendekatan stepwise AIC yang menghasilkan model terbaik berdasarkan kombinasi variabel bebas dan interaksinya. Model akhir memprediksi Wearable_Stress_Score :

Wearable_Stress_Score = 0.503911 + 0.003079(Age) -
0.008765(Social_Media_Hours) - 0.019775(Exercise_Hours) -
0.035644(Support_SystemLow) - 0.027123(Support_SystemModerate) -
0.011773(GenderF:Academic_PerformanceAverage) - 0.019288
(GenderM:Academic_PerformanceAverage) +
0.000694(GenderF:Academic_PerformanceExcellent) -
0.046803(GenderM:Academic_PerformanceExcellent) +
0.004686(GenderF:Academic_PerformanceGood) +
0.012831(GenderM:Academic_PerformanceGood) - 0.010655
(GenderF:Academic_PerformancePoor) + 0
(GenderM:Academic_PerformancePoor) +
0.004139(Social_Media_Hours:Exercise_Hours) + 0.006228
(Social_Media_Hours:Support_SystemLow) + 0.007525
(Social_Media_Hours:Support_SystemModerate) - 0.001046
(Exercise_Hours×Survey_Stress_Score) + ε

Dalam model regresi linear ini, variabel kategorik seperti Support_System, Gender, dan Academic_Performance dikodekan dalam bentuk variabel dummy. Karena metode dummy coding hanya membutuhkan n − 1 kategori, maka satu kategori dari setiap variabel dikunci sebagai kategori referensi (baseline). Kategori baseline ini tidak ditampilkan dalam model, karena koefisiennya dianggap 0 secara default dan menjadi pembanding bagi kategori lainnya.

Misalnya, dalam output model hanya muncul kategori Support_SystemLow dan Support_SystemModerate, sedangkan Support_SystemHigh tidak ditampilkan. Ini berarti High digunakan sebagai baseline. Demikian pula, GenderM:Academic_PerformancePoor tidak ditampilkan dalam model, yang

berarti bahwa kombinasi tersebut dijadikan baseline untuk interaksi Gender × Academic_Performance.

3.3.2. Uji Kelayakan Model (Uji F/fit model)

Uji kelayakan model dilakukan untuk mengetahui apakah model regresi linear yang dibentuk layak digunakan dalam memprediksi variabel respon, yaitu Wearable Stress Score. dilakukan uji kelayakan model menggunakan uji F. Uji F digunakan untuk menguji apakah setidaknya terdapat satu variabel prediktor yang secara signifikan memengaruhi variabel respon.

Hipotesis:

- H_0: Seluruh koefisien regresi dari variabel prediktor sama dengan nol (model tidak layak digunakan).
- H_1: Minimal terdapat satu koefisien regresi yang tidak nol (model layak digunakan).

Berdasarkan bentuk model yang dihasilkan, hipotesis untuk uji fit model adalah:

- $H_0 : \beta 1 = \beta 2 = \beta 3 = \cdots = \beta 16 = 0$
- H_1 : Terdapat minimal satu $\beta j \neq 0$, dimana $j = 1,2,\ldots,16$

Keputusan dalam uji kelayakan model regresi linear ditentukan dengan membandingkan nilai *p-value* hasil uji F dengan tingkat signifikansi (α) yang ditetapkan, biasanya sebesar 5%. Jika p-value < α (p-value < 0.05), maka H_0 ditolak, yang berarti model regresi secara keseluruhan layak digunakan karena terdapat minimal satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon. Sebaliknya, jika p-value ≥ α (p-value ≥ 0.05), maka H_0 gagal ditolak, yang mengindikasikan bahwa model tidak layak digunakan karena seluruh koefisien variabel prediktor tidak berbeda secara signifikan dari nol.

Berikut adalah output berdasarkan summary(model) yang dihasilkan

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.52338 -0.25114  0.00134  0.25041  0.56078

Coefficients: (1 not defined because of singularities)
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           0.503911   0.043164  11.674  < 2e-16 ***
Age                                   0.003079   0.002389   1.289  0.19744
Social_Media_Hours                   -0.008765   0.003545  -2.473  0.01345 *
Exercise_Hours                       -0.019775   0.010502  -1.883  0.05977 .
Support_SystemLow                    -0.035644   0.020125  -1.771  0.07660 .
Support_SystemModerate               -0.027123   0.019891  -1.364  0.17277
GenderF:Academic_PerformanceAverage  -0.011773   0.016188  -0.727  0.46711
GenderM:Academic_PerformanceAverage  -0.019288   0.016175  -1.192  0.23314
GenderF:Academic_PerformanceExcellent 0.000694   0.016216   0.043  0.96587
GenderM:Academic_PerformanceExcellent -0.046803  0.016379  -2.858  0.00429 **
GenderF:Academic_PerformanceGood     -0.004686   0.015832  -0.296  0.76723
GenderM:Academic_PerformanceGood      0.012831   0.016310   0.787  0.43149
GenderF:Academic_PerformancePoor     -0.010655   0.016105  -0.662  0.50825
GenderM:Academic_PerformancePoor             NA         NA      NA       NA
Social_Media_Hours:Exercise_Hours     0.004139   0.001638   2.526  0.01155 *
Social_Media_Hours:Support_SystemLow  0.006228   0.003536   1.761  0.07829 .
Social_Media_Hours:Support_SystemModerate 0.007525 0.003502  2.149  0.03170 *
Exercise_Hours:Survey_Stress_Score   -0.001046   0.001672  -0.626  0.53163
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2893 on 4983 degrees of freedom
Multiple R-squared:  0.006534,  Adjusted R-squared:  0.003344
F-statistic: 2.048 on 16 and 4983 DF,  p-value: 0.008064
```

Berdasarkan output model, diperoleh nilai statistik F sebesar 2.048 dengan p-value sebesar 0.00806 (< 0.05). Hal ini menunjukkan bahwa model secara keseluruhan signifikan pada tingkat signifikansi 5%, sehingga dapat disimpulkan bahwa terdapat minimal satu variabel bebas yang berpengaruh signifikan terhadap *Wearable Stress Score*.

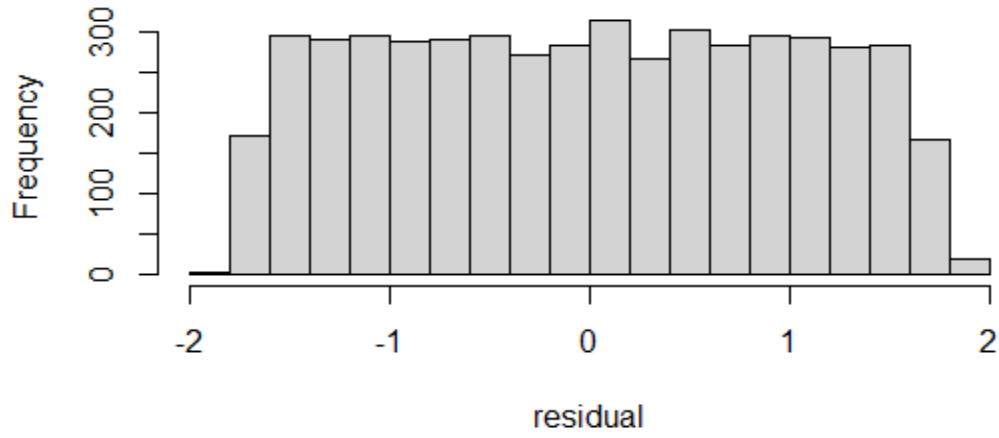
3.3.3. Koefisien Determinasi (R-squared dan Adjusted R-squared)

Nilai koefisien determinasi (R^2) sebesar 0.0065 dan Adjusted R^2 sebesar 0.0033 menunjukkan bahwa hanya sekitar 0.65% variasi Wearable Stress Score yang dapat dijelaskan oleh variabel-variabel dalam model. Meskipun model signifikan secara statistik, namun nilai R^2 yang sangat kecil menunjukkan bahwa model hanya memiliki kemampuan prediksi yang rendah, kemungkinan disebabkan oleh kompleksitas faktor-faktor psikologis yang tidak seluruhnya tercakup dalam variabel yang diamati.

3.3.4. Asumsi Uji

1. Uji Normalitas Residual

   Uji normalitas residual dilakukan untuk memastikan terpenuhinya asumsi normalitas dalam regresi linear klasik, dengan menggunakan histogram, Q-Q plot, serta uji statistik Shapiro-Wilk, Kolmogorov-Smirnov, Cramér-von Mises, dan Anderson-Darling.

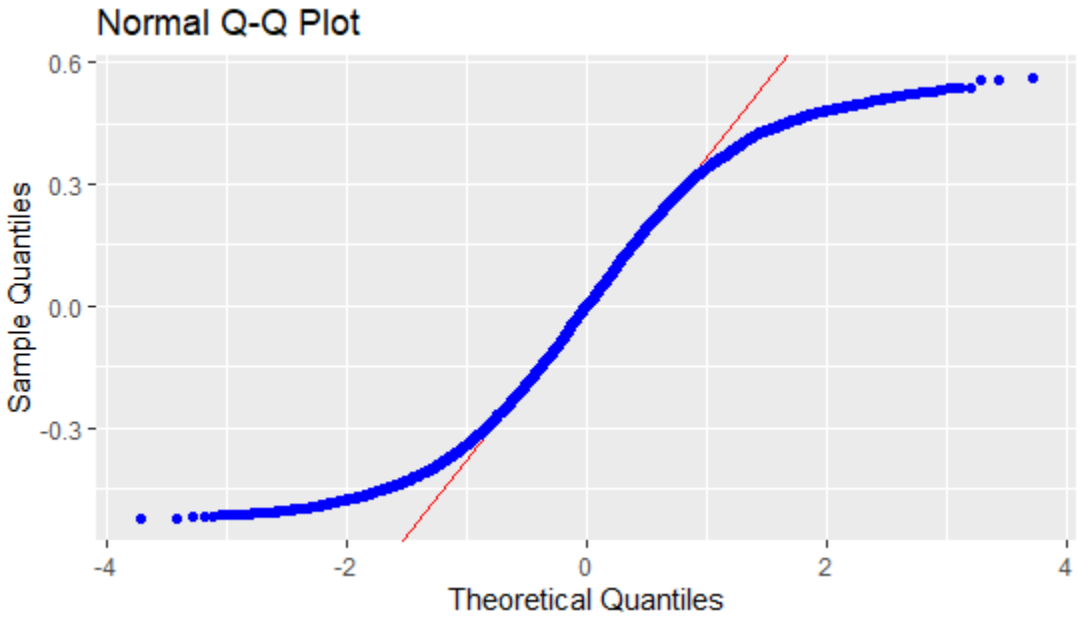


Histogram residual pada gambar menunjukkan pola distribusi yang hampir simetris dan merata di sekitar nilai nol, dengan frekuensi yang hampir sama di bagian kiri dan kanan. Pola tersebut menunjukkan bahwa residual memiliki sebaran yang mendekati distribusi normal. Hal ini mendukung asumsi normalitas yang merupakan salah satu syarat utama dalam model regresi linear klasik, sehingga model yang digunakan dapat dikatakan sudah cukup memenuhi asumsi normalitas residual. Meski demikian, agar lebih memastikan, sebaiknya juga dilakukan uji statistik normalitas seperti Shapiro-Wilk dan Kolmogorov-Smirnov.

```
> shapiro.test(resid1)

        Shapiro-Wilk normality test

data:  resid1
W = 0.95795, p-value < 2.2e-16

> ks.test(resid2, "pnorm", mean = 0, sd = 1)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  resid2
D = 0.059338, p-value = 1.022e-15
alternative hypothesis: two-sided
```

Berdasarkan hasil uji Shapiro-Wilk, diperoleh nilai statistik W = 0. 95795 dengan p-value < 2.2e-16. Nilai p yang sangat kecil (< 0.05) menunjukkan bahwa residual tidak memiliki distribusi normal secara signifikan pada tingkat kepercayaan 95%.

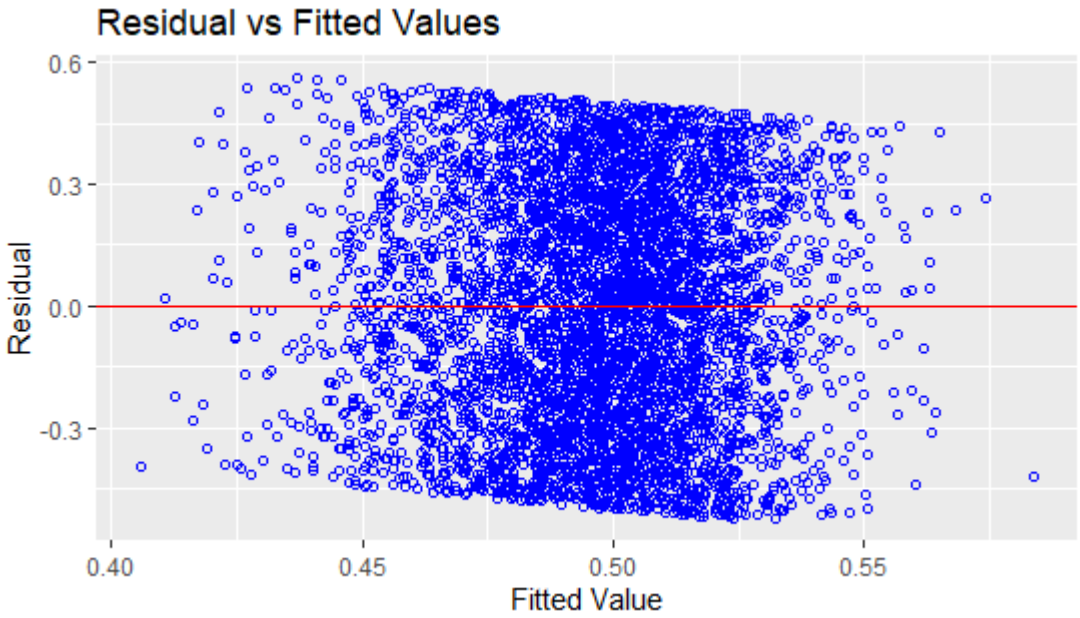
Di sisi lain, hasil uji Kolmogorov-Smirnov menunjukkan nilai statistik D = 0.059338 dan p-value = 1.022e-15, yang juga jauh di bawah 0.05. Hal ini kembali mengindikasikan bahwa hipotesis nol, yaitu residual berdistribusi normal, ditolak.



Gambar Q-Q plot di atas menunjukkan bagaimana distribusi residual dibandingkan dengan kuantil dari distribusi normal. Titik-titik biru terlihat menyebar mengikuti bentuk kurva S, dan menyimpang dari garis merah diagonal di bagian ujung-ujung grafik. Hal ini menunjukkan bahwa residual tidak sepenuhnya sesuai dengan distribusi normal, terutama pada nilai-nilai residual yang ekstrem. Meskipun bagian tengah dari titik-titik tersebut cenderung mendekati garis diagonal, penyimpangan di bagian ujung menandakan adanya ketidaksesuaian terhadap asumsi normalitas. Dengan demikian, berdasarkan Q-Q plot ini, asumsi normalitas residual tidak sepenuhnya terpenuhi.

2. Uji Homoskedastisitas

Untuk memastikan model regresi linear yang digunakan valid, salah satu asumsi klasik yang harus diuji adalah asumsi homoskedastisitas. Asumsi ini menyatakan bahwa varians dari error atau residual harus tetap konstan untuk semua nilai variabel independen. Artinya, penyebaran residual tidak boleh membentuk pola tertentu dan harus menyebar secara acak di sekitar nol. Uji homoskedastisitas penting dilakukan karena jika asumsi ini dilanggar, kondisi tersebut disebut heteroskedastisitas, yang dapat menyebabkan estimasi varians menjadi bias dan mengganggu akurasi uji statistik, seperti uji t dan uji F. Karena itu, dilakukan analisis grafik residual terhadap nilai prediksi (fitted values) untuk mengevaluasi apakah asumsi homoskedastisitas terpenuhi dalam model yang dibangun.



Plot Residual vs Fitted Values di atas menunjukkan bagaimana residual (kesalahan prediksi) terdistribusi terhadap nilai yang diprediksi oleh model regresi. Titik-titik residual tersebar secara merata di sekitar garis horizontal nol tanpa membentuk pola tertentu, yang menunjukkan bahwa asumsi homoskedastisitas, yaitu varians residual tetap konstan, sudah cukup terpenuhi. Penyebaran residual yang simetris di kedua sisi garis nol juga menunjukkan bahwa model tidak menunjukkan pola tertentu, seperti kurva atau tren, yang bisa menjadi tanda adanya kesalahan dalam spesifikasi model. Dengan demikian, grafik ini memberikan petunjuk bahwa model regresi yang digunakan telah memenuhi asumsi kenormalan dan homoskedastisitas secara visual.

```
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 11.675, df = 16, p-value = 0.766
```

Hasil Uji Breusch-Pagan yang telah distudentisasi menunjukkan nilai statistik BP sebesar 11.675 dengan derajat kebebasan (df) sebesar 16 dan p-value sebesar 0.766. Karena nilai p-value ini jauh lebih besar dari tingkat signifikansi yang umum digunakan (0.05), maka tidak ada cukup bukti untuk menolak hipotesis nol. Hipotesis nol dalam uji Breusch-Pagan menyatakan

bahwa varians residual adalah tetap atau konstan, yang dikenal sebagai homoskedastisitas.

Dengan demikian, berdasarkan hasil uji ini dapat disimpulkan bahwa tidak terdapat indikasi heteroskedastisitas, sehingga asumsi homoskedastisitas dalam model regresi dianggap terpenuhi.

3. Uji Independensi

Untuk mengetahui apakah ada perbedaan yang berarti antara dua kelompok yang tidak terkait, digunakan uji independent (uji t sampel independen). Uji ini bertujuan membandingkan rata-rata dua kelompok yang saling lepas untuk menentukan apakah perbedaan tersebut memiliki arti secara statistik. Analisis ini sangat penting untuk melihat dampak dari suatu variabel kategorik terhadap variabel numerik dalam penelitian.

```
> #Uji Independent
> durbinWatsonTest(model)
 lag Autocorrelation D-W Statistic p-value
   1     -0.01149372       2.022248   0.414
 Alternative hypothesis: rho != 0
```

Nilai Durbin-Watson (DW) mendekati 2, yaitu sebesar 2.022, menunjukkan bahwa tidak ada autokorelasi, terutama autokorelasi orde pertama, dalam residual model tersebut. Selain itu, p-value sebesar 0.414 lebih besar dari 0.05, sehingga kita tidak menolak hipotesis nol. Hal ini berarti tidak ada bukti yang cukup untuk menyatakan adanya autokorelasi dalam residual model.

Model memenuhi asumsi independensi residual. Residual tidak menunjukkan pola berurutan yang signifikan secara statistik, sehingga model regresi dapat dianggap memenuhi asumsi independensi.

3.3.5. Transformasi

Karena hasil uji menunjukkan bahwa asumsi normalitas residual belum sepenuhnya terpenuhi, dilakukan transformasi Box-Cox pada variabel dependen untuk memperbaiki distribusi residual. Setelah transformasi, model regresi dibentuk kembali dengan variabel-variabel prediktor yang sama guna mengevaluasi peningkatan kualitas model.

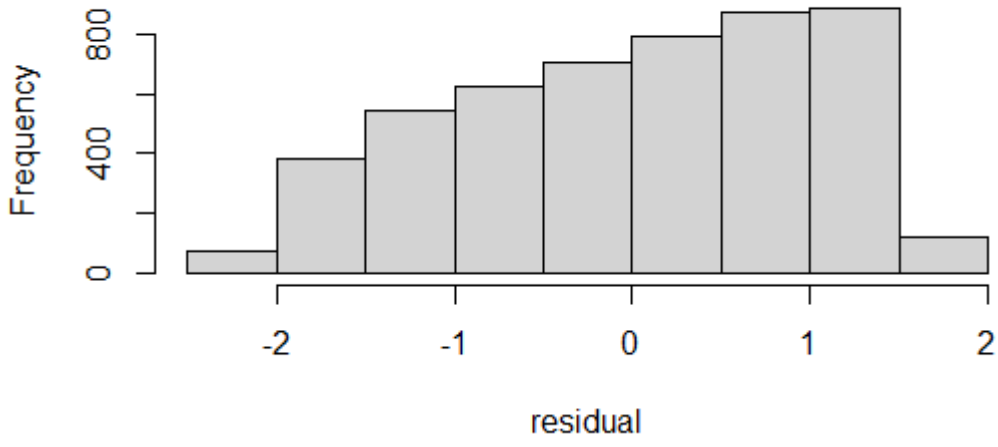
```
> # Transformasi Box-Cox di R
> pt <- powerTransform(model)
> summary(pt)
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.6976         0.7       0.6648       0.7304

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df       pval
LR test, lambda = (0) 2772.293  1 < 2.22e-16

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 285.1723  1 < 2.22e-16
```

Berdasarkan hasil transformasi Box-Cox, diperoleh nilai lambda sebesar 0,6976, yang berarti bahwa data *Wearable_Stress_Score* membutuhkan transformasi untuk mendekati distribusi normal. Hal ini diperkuat oleh hasil likelihood ratio test, yang menunjukkan bahwa baik asumsi tanpa transformasi ($\lambda = 1$) maupun log transformasi ($\lambda = 0$) ditolak secara signifikan (p-value < 2.22e-16), menandakan bahwa nilai lambda optimal berada di antara keduanya.

Diagnosis Model:

- Uji Normalitas Residual

  Uji normalitas residual dilakukan untuk memastikan terpenuhinya asumsi normalitas dalam regresi linear klasik, dengan menggunakan histogram, Q-Q plot, serta uji statistik Shapiro-Wilk, Kolmogorov-Smirnov, Cramér-von Mises, dan Anderson-Darling.



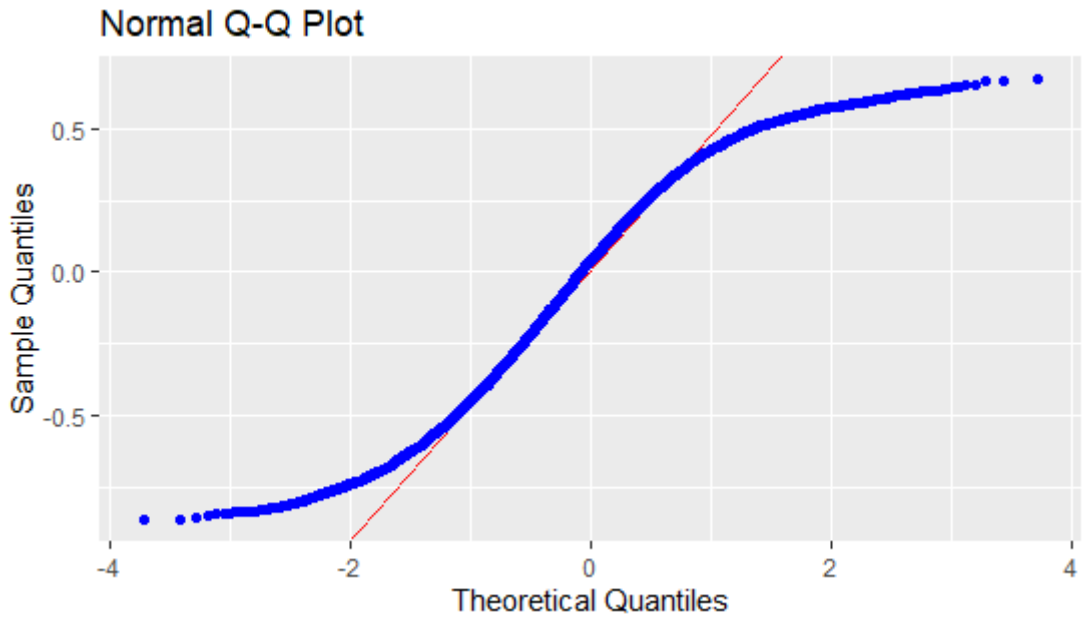
Histogram residual pada gambar menunjukkan distribusi yang tidak simetris dan condong ke kiri (left-skewed), artinya sebagian besar nilai

residual adalah positif. Pola ini menunjukkan bahwa asumsi normalitas dari residual belum terpenuhi dengan baik, yang merupakan salah satu syarat utama dalam regresi linear klasik. Ketidaksimetrisan ini bisa menunjukkan adanya bias dalam model atau mungkin perlu dilakukan transformasi tambahan pada variabel respons agar distribusi kesalahan mendekati distribusi normal. Oleh karena itu, hasil dari model regresi perlu diinterpretasikan dengan hati-hati, atau pertimbangkan perbaikan model. Sebaiknya juga dilakukan uji normalitas statistik, seperti uji Shapiro-Wilk dan Kolmogorov-Smirnov, untuk memastikan keakuratan asumsi tersebut.

```
> shapiro.test(resid1)

        Shapiro-Wilk normality test

data:  resid1
W = 0.95993, p-value < 2.2e-16

> ks.test(resid2, "pnorm", mean = 0, sd = 1)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  resid2
D = 0.067837, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Berdasarkan hasil uji Shapiro-Wilk, diperoleh nilai statistik W = 0. 95993 dengan p-value < 2.2e-16. Nilai p yang sangat kecil (< 0. 05) menunjukkan bahwa residual tidak memiliki distribusi normal secara signifikan pada tingkat kepercayaan 95%.

Di sisi lain, hasil uji Kolmogorov-Smirnov menunjukkan nilai statistik D = 0.067837 dan p-value < 2.2e-16, yang juga jauh di bawah 0.05. Hal ini kembali mengindikasikan bahwa hipotesis nol, yaitu residual berdistribusi normal, ditolak.



Gambar Q-Q plot di atas menunjukkan bagaimana distribusi residual dibandingkan dengan kuantil dari distribusi normal. Titik-titik biru terlihat menyebar mengikuti bentuk kurva S, dan menyimpang dari garis merah diagonal di bagian ujung-ujung grafik. Hal ini menunjukkan bahwa residual tidak sepenuhnya sesuai dengan distribusi normal, terutama pada nilai-nilai residual yang ekstrem. Meskipun bagian tengah dari titik-titik tersebut cenderung mendekati garis diagonal, penyimpangan di bagian ujung menandakan adanya ketidaksesuaian terhadap asumsi normalitas. Dengan demikian, berdasarkan Q-Q plot ini, asumsi normalitas residual tidak sepenuhnya terpenuhi.

Meskipun telah dilakukan transformasi Box-Cox terhadap variabel dependen untuk meningkatkan distribusi residual, hasil pemeriksaan normalitas masih menunjukkan bahwa asumsi normalitas residual belum terpenuhi.

```
> shapiro.test(e_log)

        Shapiro-Wilk normality test

data:  e_log
W = 0.81469, p-value < 2.2e-16

> # Uji normalitas Kolmogorov-Smirnov
> ks.test(e_log, "pnorm", mean = 0, sd = 1)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  e_log
D = 0.13916, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
> shapiro.test(e_inv_y)

        Shapiro-Wilk normality test

data:  e_inv_y
W = 0.045955, p-value < 2.2e-16

> # Uji normalitas Kolmogorov-Smirnov
> ks.test(e_inv_y, "pnorm", mean = 0, sd = 1)

        Asymptotic one-sample Kolmogorov-Smirnov t

data:  e_inv_y
D = 0.41922, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
> shapiro.test(e_akar)

        Shapiro-Wilk normality test

data:  e_akar
W = 0.95016, p-value < 2.2e-16

> # Uji normalitas Kolmogorov-Smirnov
> ks.test(e_akar, "pnorm", mean = 0, sd = 1)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  e_akar
D = 0.077865, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
> # Cek normalitas
> e_inv_sqrt <- rstandard(model_inv_sqrt)
> shapiro.test(e_inv_sqrt)

        Shapiro-Wilk normality test

data:  e_inv_sqrt
W = 0.25055, p-value < 2.2e-16

> # Uji normalitas Kolmogorov-Smirnov
> ks.test(e_inv_sqrt, "pnorm", mean = 0, sd = 1)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  e_inv_sqrt
D = 0.32847, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Setelah transformasi Box-Cox tidak berhasil memenuhi asumsi normalitas, dilakukan pula berbagai transformasi alternatif terhadap variabel dependen, yaitu log(y), 1/y, √y, dan 1/√y. Namun, hasil evaluasi menunjukkan bahwa seluruh bentuk transformasi tersebut juga tidak berhasil memperbaiki distribusi residual. Hal ini ditunjukkan oleh *Q-Q plot* yang masih menyimpang dari garis normal dan hasil uji statistik normalitas (seperti Shapiro-Wilk dan Kolmogorov-Smirnov) yang tetap menghasilkan p-value < 0,05, sehingga asumsi normalitas residual tetap tidak terpenuhi. Dengan demikian, permasalahan normalitas tidak dapat diatasi melalui transformasi terhadap variabel dependen.

3.4 Analisis Signifikansi Variabel

3.4.1. Uji Parsial (Uji t)

Uji parsial atau uji t digunakan untuk mengetahui dampak dari setiap variabel independen secara terpisah terhadap variabel dependen dalam model regresi. Dengan kata lain, tujuan uji ini adalah untuk mengecek signifikansi dari koefisien regresi setiap variabel prediktor, dengan menganggap variabel lainnya tetap tidak berubah. Uji dilakukan dengan membandingkan nilai p-value masing-masing koefisien terhadap tingkat signifikansi (α), yang biasanya ditentukan sebesar 0,05. Jika nilai p-value kurang dari α, maka dapat disimpulkan bahwa variabel tersebut secara statistik memiliki pengaruh yang signifikan terhadap variabel dependen. Berikut adalah uji parsial pada penelitian ini:

```
Coefficients: (1 not defined because of singularities)
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             0.503911   0.043164  11.674  < 2e-16
Age                                     0.003079   0.002389   1.289  0.19744
Social_Media_Hours                     -0.008765   0.003545  -2.473  0.01345
Exercise_Hours                         -0.019775   0.010502  -1.883  0.05977
Support_SystemLow                      -0.035644   0.020125  -1.771  0.07660
Support_SystemModerate                 -0.027123   0.019891  -1.364  0.17277
GenderF:Academic_PerformanceAverage    -0.011773   0.016188  -0.727  0.46711
GenderM:Academic_PerformanceAverage    -0.019288   0.016175  -1.192  0.23314
GenderF:Academic_PerformanceExcellent   0.000694   0.016216   0.043  0.96587
GenderM:Academic_PerformanceExcellent  -0.046803   0.016379  -2.858  0.00429
GenderF:Academic_PerformanceGood       -0.004686   0.015832  -0.296  0.76723
GenderM:Academic_PerformanceGood        0.012831   0.016310   0.787  0.43149
GenderF:Academic_PerformancePoor       -0.010655   0.016105  -0.662  0.50825
GenderM:Academic_PerformancePoor              NA         NA      NA       NA
Social_Media_Hours:Exercise_Hours       0.004139   0.001638   2.526  0.01155
Social_Media_Hours:Support_SystemLow    0.006228   0.003536   1.761  0.07829
Social_Media_Hours:Support_SystemModerate 0.007525 0.003502   2.149  0.03170
Exercise_Hours:Survey_Stress_Score     -0.001046   0.001672  -0.626  0.53163
```

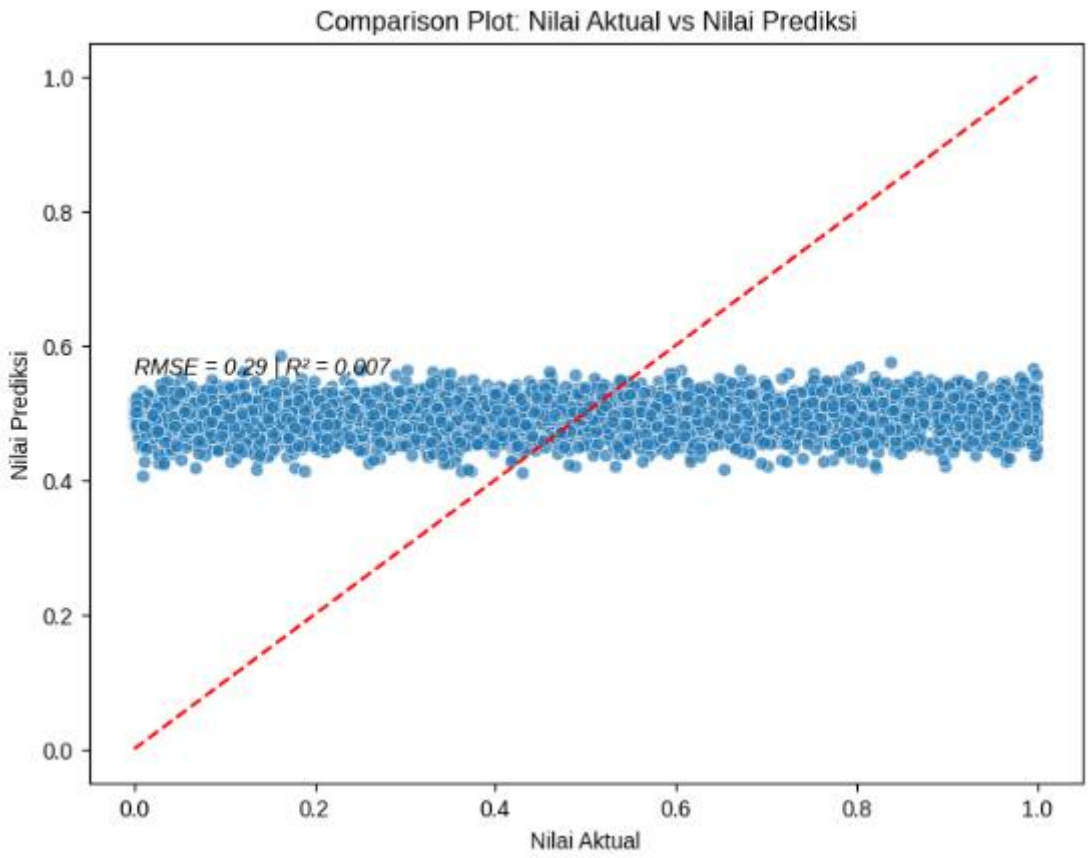
Untuk $\boldsymbol{\beta}0$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value < 0.05, maka $H0$ ditolak.	Untuk $\boldsymbol{\beta}1$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal tolak $H0$.
Untuk $\boldsymbol{\beta}2$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value < 0.05, maka $H0$ ditolak.	Untuk $\boldsymbol{\beta}3$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal tolak $H0$.
Untuk $\boldsymbol{\beta}4$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal	Untuk $\boldsymbol{\beta}5$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal

tolak $H0$.	tolak $H0$.
Untuk $\boldsymbol{\beta}6$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal tolak $H0$.	Untuk $\boldsymbol{\beta}7$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal tolak $H0$.
Untuk $\boldsymbol{\beta}8$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal tolak $H0$.	Untuk $\boldsymbol{\beta}9$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value < 0.05, maka $H0$ ditolak.
Untuk $\boldsymbol{\beta}10$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal tolak $H0$.	Untuk $\boldsymbol{\beta}11$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal tolak $H0$.
Untuk $\boldsymbol{\beta}12$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal tolak $H0$.	Untuk $\boldsymbol{\beta}13$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value tidak tersedia (NA), maka tidak dapat dilakukan pengujian hipotesis terhadap parameter tersebut. Hal ini menunjukkan adanya masalah dalam model, seperti multikolinearitas atau struktur data yang tidak memadai, yang perlu diperiksa dan disesuaikan

	sebelum interpretasi lebih lanjut dapat dilakukan.
Untuk $\boldsymbol{\beta}14$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value < 0.05, maka $H0$ ditolak.	Untuk $\boldsymbol{\beta}15$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal tolak $H0$.
Untuk $\boldsymbol{\beta}16$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value < 0.05, maka $H0$ ditolak.	Untuk $\boldsymbol{\beta}17$ (Intercept): Hipotesis: $H0 : \beta 0 = 0$ $H1 : \beta 0 \neq 0$ Jika p-value > 0.05, maka gagal tolak $H0$. Jika p-value < 0.05, maka $H0$ ditolak. Kesimpulan: Karena p-value > 0.05, maka gagal tolak $H0$.

3.5 Visualisasi Hasil

Untuk mengecek bagaimana baiknya model regresi linear yang dibuat, dilakukan visualisasi perbandingan antara nilai sebenarnya dan nilai yang diprediksi dari variabel Wearable Stress Score. Tujuan dari visualisasi ini adalah untuk melihat seberapa dekat hasil prediksi dengan nilai asli. Semakin banyak titik data yang mendekati garis acuan y=x, maka semakin baik kemampuan model dalam memprediksi. Berikut ini adalah grafik yang menunjukkan perbandingan antara nilai aktual dan nilai prediksi:



Berdasarkan perbandingan antara nilai aktual dan nilai prediksi, tampak bahwa kebanyakan titik data berada jauh dari garis acuan y = x, yang mengartikan bahwa hasil prediksi tidak mendekati nilai sebenarnya. Koefisien determinasi (R^2) yang didapatkan adalah 0. 007, artinya model hanya mampu menjelaskan sekitar 0. 7% dari perubahan variabel dependen, sehingga kemampuan prediksi model sangat rendah. Selain itu, nilai RMSE sebesar 0. 29 menunjukkan adanya kesalahan prediksi yang cukup besar dalam skala data yang digunakan. Secara keseluruhan, model regresi ini belum menunjukkan hasil yang memuaskan dan membutuhkan evaluasi atau peningkatan lebih lanjut.

BAB IV

KESIMPULAN

4.1 Kesimpulan

Hasil penelitian ini menunjukkan bahwa model regresi linear berganda yang dibangun untuk memprediksi tingkat stres remaja berdasarkan dataset Mental health analysis among teenagers memiliki performa yang sangat rendah dalam menjelaskan variabilitas data. Meskipun model secara statistik signifikan (uji F, p-value < 0.05), nilai koefisien determinasi ($R^2 = 0.0065$) dan Adjusted R^2 (0.0033) mengindikasikan bahwa hanya sekitar 0.65% variasi dalam skor stres wearable yang dapat dijelaskan oleh variabel-variabel prediktor dalam model.

Beberapa variabel seperti durasi penggunaan media sosial, performa akademik, dan interaksi antara media sosial dengan olahraga memiliki pengaruh yang signifikan secara statistik, namun besarnya kontribusi terhadap total variasi stres sangat kecil. Selain itu, uji validasi silang 10-fold menghasilkan nilai R^2 yang serupa rendahnya, memperkuat kesimpulan bahwa daya prediksi model terhadap data baru masih lemah.

Meskipun semua uji asumsi regresi (normalitas, multikolinearitas, homoskedastisitas, dan independensi) sebagian besar telah terpenuhi, model ini belum mampu menangkap kompleksitas hubungan antara variabel-variabel kehidupan remaja dengan tingkat stres yang tercatat secara fisiologis oleh wearable device.

Model linier klasik memiliki keterbatasan dalam menjelaskan stres fisiologis. Hal ini mengindikasikan perlunya pendekatan alternatif, seperti model non-linear atau machine learning, yang dapat menangkap pola hubungan kompleks dan interaksi variabel yang lebih fleksibel. Variabel-variabel yang digunakan belum cukup mewakili faktor-faktor utama penyebab stres fisiologis. Diperlukan pengayaan data, seperti memasukkan aspek biologis, hormonal, atau psikososial yang lebih mendalam. Hasil ini memberikan kontribusi awal dalam pemahaman kuantitatif tentang stres remaja berbasis data wearable, sekaligus menegaskan bahwa pendekatan statistik konvensional perlu dikombinasikan dengan data dan metode yang lebih adaptif untuk menghasilkan prediksi yang akurat dan bermakna.

4.2 Saran

Berdasarkan analisis yang telah dilakukan, dapat disimpulkan bahwa meskipun model regresi linear berganda yang dibuat memenuhi hampir semua asumsi klasik dan

dibuat dengan metode yang sistematis serta validasi yang cukup baik, kemampuan model dalam memprediksi masih sangat rendah. Hal ini tampak dari nilai R kuadrat dan Adjusted R kuadrat yang sangat kecil, baik pada model awal maupun hasil validasi silang, yang menunjukkan bahwa model hanya mampu menjelaskan sebagian kecil dari variasi tingkat stres yang terukur oleh perangkat wearable. Oleh karena itu, disarankan untuk meninjau pendekatan non-linear atau metode machine learning seperti Random Forest atau Gradient Boosting, karena lebih efektif dalam menangkap pola dan hubungan antar variabel yang kompleks. Selain itu, perlu juga dilakukan eksplorasi dan penambahan variabel tambahan yang lebih relevan dan kontekstual, seperti kondisi psikologis spesifik, riwayat stres, atau data fisiologis lainnya. Strategi lain yang dapat diterapkan adalah melakukan segmentasi data berdasarkan kelompok tertentu, seperti jenis kelamin, usia, atau tingkat pendidikan, sehingga model yang dibuat lebih akurat dan spesifik sesuai dengan karakteristik populasi yang dituju.

REFERENSI

Wankhede, A. (2023). *Mental health analysis among teenagers* [Dataset]. Kaggle. https://www.kaggle.com/datasets/aniruddhawankhede/mental-heath-analysis-among-teenagers

Wijaya, M. Y. (n.d.). *Model linear 1 (Review statistika dasar)* [Modul perkuliahan]. Program Studi Matematika, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta.

Wijaya, M. Y. (n.d.). *Model linear 2 (Pengantar model linear)* [Modul perkuliahan]. Program Studi Matematika, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta.

Wijaya, M. Y. (n.d.). *Model linear 3 (Regresi Linear Sederhana)* [Modul perkuliahan]. Program Studi Matematika, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta.

Wijaya, M. Y. (n.d.). *Model linear 4 (Regresi Linear Berganda)* [Modul perkuliahan]. Program Studi Matematika, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta.

Wijaya, M. Y. (n.d.). *Model linear 5 (Seleksi Model)* [Modul perkuliahan]. Program Studi Matematika, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta.

Wijaya, M. Y. (n.d.). *Model linear 6 (Diagnosis Model)* [Modul perkuliahan]. Program Studi Matematika, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta.

Wijaya, M. Y. (n.d.). *Model linear 7 (Perbaikan & Validasi Model: Remedial Measures & Model Validation))* [Modul perkuliahan]. Program Studi Matematika, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta.

APENDIKS

Link Google Collab : https://colab.research.google.com/drive/1TLTm3Vw_E-ITIxTAVWiW6dNe1sk4m73T?usp=sharing