Diterima : 24 November 2024
Disetujui : 11 Desember 2024
Hal : 22-30

Al-Aqlu: Jurnal Matematika, Teknik dan Sains
https://jurnal.yalamqa.com/index.php/aqlu
e-ISSN : 2985-4369
Vol. 3 , No. 1, Januari 2025

PREDIKSI INDEKS PERFORMA SISWA BERDASARKAN WAKTU BELAJAR, NILAI SEBELUMNYA, KEGIATAN EKSTRAKURIKULER, WAKTU TIDUR, DAN BANYAKNYA SOAL YANG DIKERJAKAN DENGAN REGRESI LINEAR KUADRAT-TERKECIL

[Prediction of Student Performance Index Based on Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, and Sample Question Papers Practiced with Least-Squares Linear Regression]

Sapto Mukti Handoyo[1)], Mohamad Khoirun Najib[2)*]

IPB University

mkhoirun@apps.ipb.ac.id (corresponding)

ABSTRAK

Indeks performa siswa merupakan suatu ukuran yang digunakan untuk menyatakan performa siswa secara keseluruhan. Salah satu model yang dapat digunakan untuk memprediksi indeks performa siswa adalah model regresi berbasis *machine learning*. Oleh karenanya, penelitian ini bertujuan untuk menerapkan model regresi linear kuadrat-terkecil berbasis *machine learning* untuk memprediksi indeks performa berdasarkan faktor-faktor tersebut dan menginterpretasikan model tersebut. Model regresi yang digunakan tersedia pada paket pemrograman Julia yang bernama MLJ. Model ini dievaluasi berdasarkan beberapa kriteria, yaitu nilai R-*squared*, RMSE, dan MAE. Hasilnya menunjukan bahwa faktor nilai sebelumnya berpengaruh paling besar terhadap indeks performa siswa. Kemudian, nilai R-*squared* untuk data uji diperoleh sebesar 0.988, RMSE untuk data latih sebesar 0.106, RMSE untuk data uji sebesar 0.108, MAE untuk data latih sebesar 0.84, dan MAE untuk data uji sebesar 0.86. Berdasarkan hasil evaluasi, model memiliki kemampuan prediksi yang baik dengan galat rata-rata yang rendah, tidak mengalami *overfitting*, dan memiliki kemampuan generalisasi yang baik.

Kata kunci: *Machine learning; indeks performa siswa; regresi kuadrat-terkecil; regresi linear berganda.*

ABSTRACT

The students' performance index is a measure used to represent the overall performance of students. One model that can be used to predict the students' performance index is a machine learning-based regression model. Therefore, this study aims to apply a machine learning-based least-squares linear regression model to predict the performance index using these factors and interpret the model. The regression model utilized is available in the Julia programming package called MLJ. This model is evaluated based on several criteria, including R-squared, RMSE, and MAE. The results show that the previous scores have the most significant influence on the students' performance index. Furthermore, the R-squared value for the test data is 0.988, the RMSE for the training data is 0.106, the RMSE for the test data is 0.108, the MAE for the training data is 0.84, and the MAE for the test data is 0.86. Based on the evaluation results, the model has good predictive performance with low average error, does not experience overfitting, and has good generalization ability.

Keywords: *Least-square regression; machine learning; multiple linear regression; students' performance index.*

PENDAHULUAN

Machine learning merupakan cabang dari ilmu kecerdasan buatan (*articial intelligence*) yang membangun dan mengembangkan algoritme untuk mengembangkan perilaku pada komputer berdasarkan data empiris (Jalal dan Ezzedine 2019). Studi tentang teori *computational learning* dan pengenalan pola telah membuat *machine learning* berkembang. Selain itu, *machine learning* merupakan metode yang efektif untuk memprediksi sesuatu dengan menggunakan beberapa model dan algoritme tertentu (Angra dan Ahuja 2017). *Machine learning* dapat diklasifikasikan menjadi tiga jenis, yaitu *supervised*, *unsupervised*, dan *reinforcement learning*. *Supervised learning* adalah jenis machine learning yang melatih input data yang telah diberi label untuk memprediksi suatu *output* tertentu (*target*). Contoh algoritme *supervised learning* adalah klasifikasi dan regresi. Sedangkan, *unsupervised learning* tidak membutuhkan label pada dataset. *Unsupervised learning* adalah jenis *machine learning* yang menggunakan kumpulan data tanpa kejelasan peubah respons. Artinya, komputer mempelajari pola data tanpa merujuk ke suatu respons tertentu. Contoh algoritme *unsupervised learning* adalah pengelompokkan (*clustering*) dan transformasi data informatif (Haldorai, dkk. 2019, Morales dan Escalante 2022, Valkenborg, dkk. 2023). *Reinforcement learning* adalah jenis *machine learning* yang didasarkan pada pemetaan dari situasi ke tindakan untuk memaksimalkan imbalan (*reward*) dengan mencari tindakan yang menghasilkan *reward* tertinggi (Sutton 1992).

Machine learning merupakan salah satu metode yang terkenal untuk digunakan dalam analisis dan prediksi data. Berdasarkan data di situs web scopus.com, terdapat 86,295 artikel ilmiah terkait machine learning yang dilakukan pada tahun 2024. Pada tahun yang sama, artikel yang paling banyak terbit di Scopus merupakan artikel yang berada di bidang Ilmu Komputer, Keteknikan, dan Matematika. Penelitian terkait *machine learning* untuk peramalan telah banyak dilakukan oleh beberapa peneliti (Zhang, dkk. 2024, Huang, dkk. 2024, Htun, dkk. 2024). Selain itu, penggunaan *machine learning* untuk prediksi juga telah banyak dilakukan oleh beberapa peneliti (Nixon dan Gilbert 2024, Gu, dkk. 2024, Mamo, dkk. 2024).

Machine learning dapat digunakan untuk memprediksi indeks performa siswa. Indeks performa mengukur performa siswa secara keseluruhan yang menunjukkan performa akademik siswa. Penelitian terkait dengan performa siswa telah banyak dilakukan oleh beberapa peneliti, seperti klasifikasi dan prediksi performa siswa menggunakan algoritme *support vector machines*, Naive Bayes, dan algoritme lain (Pallathadka, dkk. 2023), prediksi performa akademik siswa menggunakan algoritme *decision tree regression* dan *decision tree classifier* (Hussain dan Khan 2023), prediksi performa siswa menggunakan algoritme *artificial neural network*, *decision tree*, *logistic regression*, dan Naive Bayes (Altabrawee, dkk. 2019), prediksi performa akademik siswa menggunakan algoritme *random forest*, *nearest neighbour*, *support vector machines*, *logistic regression*, Naive Bayes, dan k-*nearest neighbour* (Yağcı 2022), dan prediksi performa siswa dengan koefisien pembelajaran menggunakan algoritme regresi linear, *decision tree*, *random forest*, dan *support vector regression* (Asthana, dkk. 2023).

Indeks performa dapat dipengaruhi oleh beberapa faktor, seperti waktu belajar, waktu tidur, keikutsertaan dalam kegiatan ekstrakurikuler, dan faktor-faktor lainnya. Faktor-faktor tersebut dapat dijadikan sebagai dasar pertimbangan bagi guru dalam membuat strategi peningkatan indeks performa siswa. Oleh karena itu, diperlukan suatu model *machine learning* yang dapat digunakan untuk memprediksi indeks performa siswa berdasarkan faktor-faktor yang diduga dapat berpengaruh. Dalam penelitian ini, model *machine learning* yang digunakan adalah regresi kuadrat-terkecil. Faktor-faktor yang diduga berpengaruh terhadap indeks performa siswa adalah waktu belajar, nilai sebelumnya, keikutsertaan dalam kegiatan ekstrakurikuler, waktu tidur, dan banyaknya soal yang dikerjakan. Model regresi yang diperoleh dievaluasi menggunakan R-*squared*, *Root Mean Squared Error* (RMSE), dan *Mean Absolute Error* (MAE).

METODE PENELITIAN

Alat dan Metode

Alat yang digunakan pada penelitian ini adalah *personal computer* (PC) ASUS dengan model sistem X441B. PC yang digunakan menggunakan sistem operasi Windows 10 64-bit, prosesor AMD A9-9425 Radeon R5 2 CPU 3.1 GHz, dan RAM 4 GB. Perangkat lunak yang digunakan pada penelitian ini adalah Microsoft Excel 2019 untuk membuka data dengan format file CSV dan Julia versi 1.10.4 untuk

menganalisis data dan memodelkan data dengan regresi. Beberapa paket yang digunakan di dalam Julia dapat dilihat pada Tabel 1 berikut.

Tabel 1. Paket yang Digunakan di dalam Julia

Nama Paket	Kegunaan
DataFrames	Membuat data dalam bentuk tabel, memanipulasinya, mengelompokkannya, dan menghitung statistik deskriptif.
CSV	Membaca data dengan format file CSV.
PrettyPrinting	Menampilkan *output* yang lebih terstruktur dan mudah dibaca.
MLJ	Memilih, menggunakan, melatih, dan mengevaluasi model *machine learning*.
ScientificTypes	Menentukan tipe saintifik (ilmiah) pada data.
Statistics	Menghitung statistik dasar, seperti mean, median, simpangan baku, dan lain-lain.
Plots	Membuat grafik dan plot, seperti grafik garis, batang, plot sebar, dan lain-lain.
StatsPlots	Membuat plot yang lebih khusus dan lebih sesuai dengan kebutuhan.

Data

Data yang digunakan pada penelitian ini merupakan data sekunder tentang data performa siswa yang berisi 6 peubah, yaitu waktu belajar, nilai sebelumnya, kegiatan ekstrakurikuler, waktu tidur, banyaknya soal yang dikerjakan, dan indeks performa. Data tersebut berukuran 10000×6 dan terdiri dari 60 ribu amatan. Selain itu, data tersebut memiliki format CSV dan bersumber dari situs web Kaggle dengan link: https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression.

Regresi Linear Berganda

Penelitian ini menggunakan 5 peubah prediktor, yaitu waktu belajar, nilai sebelumnya, kegiatan ekstrakurikuler, waktu tidur, dan banyaknya soal yang dikerjakan, dan 1 peubah respon, yaitu indeks performa siswa. Model regresi yang digunakan adalah model regresi linear berganda dan memiliki bentuk sebagai berikut.

$$y = f_\beta(\boldsymbol{x}) + \varepsilon = \beta_0 + \sum_{k=1}^{5} \beta_k x_k + \varepsilon. \quad (1)$$

dengan
x_1 menyatakan waktu belajar siswa (jam),
x_2 menyatakan nilai sebelumnya,
x_3 menyatakan keikutsertaan dalam kegiatan ekstrakurikuler,
x_4 menyatakan waktu tidur siswa (jam),
x_5 menyatakan banyaknya contoh soal yang dikerjakan,
y adalah nilai indeks performa siswa,
β_k adalah nilai koefisien dari x_k, $k = 1,2,\dots,5$,
β_0 adalah nilai koefisien intersep (perpotongan), dan
$f_\beta(\boldsymbol{x})$ adalah nilai prediksi dari y.

Hal yang ingin ditentukan dari model regresi adalah menduga nilai-nilai koefisien atau parameter β_k untuk setiap $k = 0,1,\dots,5$. Salah satu pendekatan *machine learning* yang dapat digunakan untuk menduga koefisien tersebut adalah regresi kuadrat-terkecil.

Model regresi kuadrat-terkecil memilih nilai koefisien yang terbaik dengan meminimumkan nilai rata-rata dari *squared loss* atau *mean squared error* (MSE) (Heath 2002, Johnson dan Faunt 1992, Popiński 1993) dan dapat dituliskan sebagai berikut.

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{N_l} \sum_{i=1}^{N_l} \left(f_\beta\left(x^{(i)}\right) - y^{(i)}\right)^2. \quad (2)$$

dengan N_l adalah banyaknya data latih, $\hat{\beta}$ nilai dugaan bagi koefisien β, $y^{(i)}$ dan $f_\beta\left(x^{(i)}\right)$ masing-masing berturut-turut adalah nilai aktual dan prediksi dari indeks performa siswa ke-i.

Evaluasi Model

Model *machine learning* yang digunakan pada data uji dievaluasi dengan beberapa metrik. Metrik yang digunakan adalah *R-squared*, *Root Mean Squared Error* (RMSE), dan *Mean Absolute Error* (MAE) yang dapat dituliskan sebagai berikut.

$$R^2 = 1 - SSR/SST \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N_u}\sum_{i=1}^{N_u}\left(f_{\hat{\beta}}(x^{(i)}) - y^{(i)}\right)^2}, \tag{4}$$

$$MAE = \frac{1}{N_u}\sum_{i=1}^{N_u}\left|f_{\hat{\beta}}\left(x^{(i)}\right) - y^{(i)}\right|. \tag{5}$$

dengan SSR adalah jumlah kuadrat residual, SST adalah jumlah kuadrat total, N_u adalah banyaknya data uji, $y^{(i)}$ dan $f_{\hat{\beta}}\left(x^{(i)}\right)$ masing-masing berturut-turut adalah nilai aktual dan prediksi indeks performa siswa ke-i berdasarkan model regresi dengan nilai koefisien atau parameter $\hat{\beta}$.

Tahapan

Berikut ini merupakan langkah-langkah yang dilakukan dalam mengerjakan makalah ini.

1) Identifikasi Masalah

   Penelitian dimulai membaca referensi mengenai indeks performa siswa. Selain itu, identifikasi mengenai peubah-peubah yang berhubungan terhadap indeks performa siswa juga dilakukan untuk melihat kemungkinan pengaruhnya terhadap indeks performa siswa, sehingga indeks performa siswa dapat diprediksi.
2) Studi Literatur

   Tahap awal dari studi literatur adalah membaca referensi mengenai model-model prediksi yang mungkin sesuai permasalahan yang ingin dibahas. Dalam hal ini, model yang mungkin sesuai adalah model regresi. Kemudian, dilakukan studi literatur mengenai konsep matematis dari regresi linear berganda dengan jenis regresi yang digunakan adalah regresi kuadrat-terkecil. Selain itu, studi literatur mengenai penggunaan perangkat lunak Julia dalam melakukan analisis data dan prediksi data dengan *machine learning* juga dilakukan.
3) Pengumpulan Data

   Dalam penelitian ini, data yang dikumpulkan merupakan data sekunder yang diperoleh dari situs web Kaggle. Data tersebut terdiri dari waktu belajar, nilai sebelumnya, kegiatan ekstrakurikuler, waktu tidur, banyaknya soal yang dikerjakan, dan indeks performa.
4) Pemrosesan Data

   Data yang telah dikumpulkan diproses untuk melihat kualitas dan kesesuaian data dengan model prediksi yang akan digunakan. Dalam hal ini, dilakukan pemeriksaan dan penanganan mengenai jenis data, data yang hilang, penghapusan data, dan transformasi data jika diperlukan.
5) Analisis Data Eksploratif

   Analisis data eksploratif dilakukan untuk mengetahui karakteristik data yang digunakan. Analisis dilakukan dengan melihat kecenderungan data, sebaran data, dan korelasi antar peubah.
6) Pemodelan Regresi Kuadrat-Terkecil

   Data dipartisi menjadi dua bagian, yaitu data latih dan data uji dengan perbandingan 80:20. Kemudian, model prediksi untuk memprediksi peubah respon (indeks performa) berdasarkan peubah prediktor (waktu belajar, nilai sebelumnya, kegiatan ekstrakurikuler, waktu tidur, dan banyaknya soal yang dikerjakan) dimodelkan menggunakan regresi kuadrat-terkecil. Dalam hal ini, model tersebut dibangun sehingga diperoleh model dengan nilai dugaan parameter terbaik.
7) Evaluasi dan Interpretasi Model

   Kinerja dari model yang diperoleh diukur menggunakan beberapa metrik, yaitu *R-squared*, RMSE, dan MAE. Kemudian, interpretasi model dilakukan dengan melihat nilai koefisien regresi untuk melihat pengaruh masing-masing peubah terhadap indeks performa siswa.

8) Kesimpulan

Kesimpulan diperoleh dari hasil penelitian yang telah dilakukan secara umum serta dapat menjawab permasalahan dan tujuan penelitian yang ingin diselesaikan.

HASIL DAN PEMBAHASAN

Bagian ini menjelaskan ringkasan data, transformasi data, dan hasil pendugaan parameter dari model regresi kuadrat-terkecil. Kemudian, model yang diperoleh dievaluasi menggunakan metrik RMSE dan MAE. Model tersebut diinterpretasikan untuk melihat pengaruh setiap peubah terhadap indeks performa siswa.

Ringkasan Data

Subbagian ini akan menunjukkan karakteristik data yang digunakan. Berikut ini disajikan empat data awal yang digunakan pada penelitian ini yang dapat dilihat pada Tabel 2.

Tabel 2. Empat Data Awal Penelitian

Peubah	Nilai ke-			
	1	**2**	**3**	**4**
HoursStudied	7	4	8	5
PreviousScores	99	83	51	52
ExtracurricularActivities	*Yes*	*No*	*Yes*	*Yes*
SleepHours	9	4	7	5
SampleQuestionPapersPracticed	1	2	2	2
PerformanceIndex	91.0	65.0	45.0	36.0

Berdasarkan Tabel 2, dapat dilihat bahwa terdapat peubah kategorik, yaitu *ExtracurricularActivities* karena nilainya berupa nilai kategorik (*Yes* atau *No*), sedangkan peubah lainnya merupakan peubah numerik. Selain itu, peubah numerik *PerformanceIndex* berjenis Float64, sedangkan peubah numerik lainnya berjenis Int64.

Model regresi hanya dapat diterapkan pada data yang jenisnya numerik dan kontinu. Oleh karena itu, diperlukan pemeriksaan terhadap tipe saintifik dari data yang digunakan. Tipe saintifik dari data tersebut dapat dilihat pada Tabel 3 berikut ini.

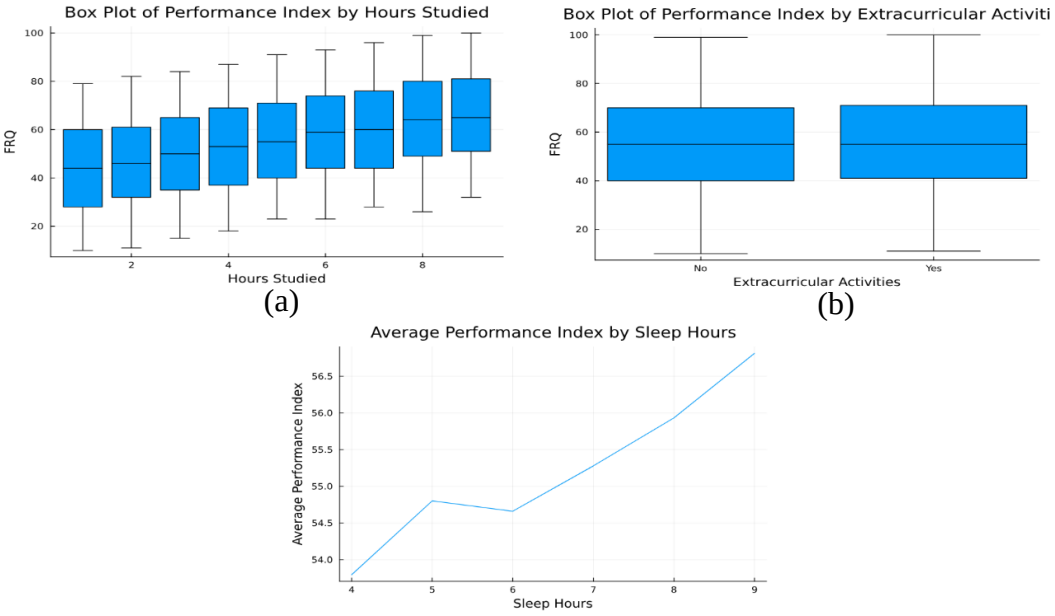
Tabel 3. Tipe Saintifik Data

Peubah	Tipe Saintifik
HoursStudied	*Count*
PreviousScores	*Count*
ExtracurricularActivities	*Textual*
SleepHours	*Count*
SampleQuestionPapersPracticed	*Count*
PerformanceIndex	*Continuous*

Berdasarkan Tabel 3, dapat dilihat bahwa peubah kategorik, yaitu *ExtracurricularActivities* memiliki tipe saintifik *Textual*. Selain itu, peubah numerik *PerformanceIndex* memiliki tipe saintifik *Continuous*, sedangkan peubah numerik lainnya memiliki tipe saintifik *Count*. Dalam hal ini, peubah kategorik harus diubah jenisnya menjadi numerik dan tipe saintifik seluruh peubah numerik harus diubah menjadi *Continuous* sebelum regresi dilakukan.

Data yang dikumpulkan memiliki duplikat antara satu baris dengan baris lainnya. Banyaknya data yang terduplikasi adalah 127. Selain itu, data tersebut tidak memiliki nilai yang kosong atau hilang. Dalam hal ini, data duplikat harus dihapus sebelum regresi dilakukan agar dapat diterapkan ke dalam model regresi.

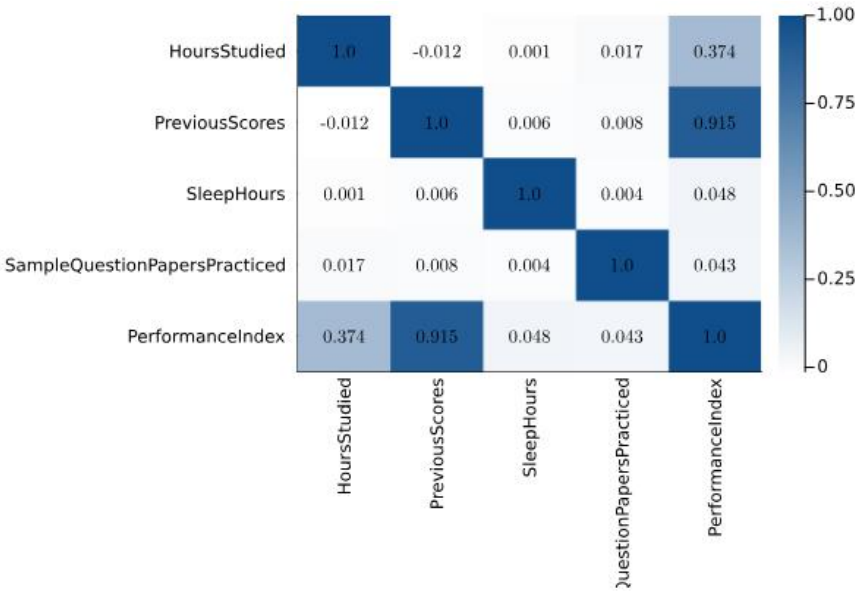
Ringkasan singkat secara visual tentang data yang digunakan dapat dibuat menggunakan *boxplot* dan *line plot*. Ringkasan singkat tersebut dapat dilihat pada Gambar 2 berikut.



Gambar 1. Visualisasi hubungan antar peubah: (a) *boxplot* indeks performa terhadap waktu belajar (b) *boxplot* indeks performa terhadap kegiatan ekstrakurikuler (c) *line plot* indeks performa rata-rata terhadap waktu tidur

Berdasarkan Gambar 1, diperoleh bahwa semakin lama waktu belajar siswa, indeks performa siswa semakin meningkat. Kemudian, partisipasi siswa dalam kegiatan ekstrakurikuler sedikit berperan dalam meningkatkan indeks performa. Selain itu, semakin lama waktu tidur siswa, indeks performa rata-ratanya semakin meningkat.

Hubungan antara peubah prediktor dengan peubah respon dapat dilihat menggunakan *heatmap* yang dapat menunjukkan korelasi antara dua peubah. *Heatmap* tersebut dapat dilihat pada Gambar 1 berikut.



Gambar 2. *Heatmap* antara peubah prediktor dan respon

Berdasarkan Gambar 2, dapat diperoleh bahwa korelasi antara peubah prediktor dan peubah respon semuanya bernilai positif. Hal ini menunjukkan bahwa peningkatan nilai peubah prediktor cenderung akan meningkatkan nilai peubah respon. Korelasi yang paling kuat terjadi antara peubah prediktor *PreviousScores* dan peubah respon *PerformanceIndex* karena nilainya mendekati 1. Korelasi yang paling

lemah terjadi antara peubah prediktor *SampleQuestionPapersPracticed* dan peubah respon *PerformanceIndex* karena nilainya mendekati 0.

Transformasi Data

Seperti yang telah disebutkan pada subbagian sebelumnya, model regresi hanya dapat diterapkan pada data yang numerik, kontinu, dan tidak mengandung duplikat. Oleh karena itu, sebelum melakukan pendugaan parameter, data yang duplikat dihapus terlebih dahulu menggunakan fungsi *unique(dataframe)* di Julia. Ukuran data setelah data duplikat dihapus adalah 9873 × 6. Berikut ini disajikan empat data awal setelah dilakukan konversi jenis data pada Tabel 4.

Tabel 4. Empat Data Awal Penelitian Setelah Jenis Data Dikonversi

Peubah	Nilai ke-			
	1	**2**	**3**	**4**
HoursStudied	7.0	4.0	8.0	5.0
PreviousScores	99.0	83.0	51.0	52.0
ExtracurricularActivities	1.0	0.0	1.0	1.0
SleepHours	9.0	4.0	7.0	5.0
SampleQuestionPapersPracticed	1.0	2.0	2.0	2.0
PerformanceIndex	91.0	65.0	45.0	36.0

Berdasarkan Tabel 4, data kategorik *ExtracurricularActivities* dikonversi menjadi data numerik yang nilainya 0 jika nilai kategoriknya *No* dan 1 jika nilai kategoriknya *Yes*. Selanjutnya, semua data numerik yang berjenis Int64 dikonversi menjadi Float64 sehingga diperoleh tipe saintifik semua peubahnya adalah *Continuous*.

Pengaruh yang diberikan dari peubah prediktor terhadap peubah respon dalam model regresi dapat dilihat dari koefisien regresinya. Agar dapat melihat pengaruhnya dengan benar, data tersebut diskalakan ulang dengan metode normalisasi karena setiap peubah memiliki skala yang cukup berbeda. Normalisasi dilakukan untuk seluruh peubah kecuali peubah *ExtracurricularActivities*.

Pendugaan Parameter Model Regresi

Dengan menggunakan data latih, koefisien regresi diduga menggunakan regresi kuadrat-terkecil. Koefisien ini dapat dihitung menggunakan matriks. Misalkan X merupakan matriks yang mengandung himpunan peubah prediktor dan $\boldsymbol{\beta}$ merupakan vektor yang mengandung koefisien regresi sebagai berikut.

$$X = [\mathbf{1} \quad \boldsymbol{X}_1 \quad \boldsymbol{X}_2 \quad \boldsymbol{X}_3 \quad \boldsymbol{X}_4 \quad \boldsymbol{X}_5], \tag{6}$$

$$\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4 \quad \beta_5]^T. \tag{7}$$

dengan $\boldsymbol{X}_k$ merupakan vektor kolom peubah prediktor ke-k pada data latih, β_k merupakan koefisien regresi dari peubah prediktor ke-k untuk $k = 1,2,\dots,5$, dan β_0 merupakan koefisien intersep. Misalkan $\boldsymbol{Y}$ merupakan vektor kolom yang mengandung peubah respon *PerformanceIndex* pada data latih. Maka, nilai dugaan koefisien $\widehat{\boldsymbol{\beta}}$ sebagai penduga untuk $\boldsymbol{\beta}$ adalah

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1}(X^T \boldsymbol{Y}). \tag{8}$$

Berdasarkan data latih, model regresi kuadrat-terkecil diperoleh sebagai berikut.

$$\hat{y} = f_{\hat{\beta}}(\boldsymbol{x}) = -0.017 + 0.384x_1 + 0.918x_2 + 0.042x_3 + 0.029x_4 + 0.33x_5. \tag{9}$$

Model tersebut dapat interpretasi sebagai berikut.

1) kenaikan satu standar deviasi pada x_1 (waktu belajar siswa) menyebabkan peningkatan sebesar 0.384 standar deviasi pada $\hat{y}$ (indeks performa siswa),

2) kenaikan satu standar deviasi pada x_2 (nilai sebelumnya) menyebabkan peningkatan sebesar 0.918 standar deviasi pada $\hat{y}$ (indeks performa siswa),
3) kenaikan satu standar deviasi pada x_3 (keikutsertaan dalam kegiatan ekstrakurikuler) menyebabkan peningkatan sebesar 0.042 standar deviasi pada $\hat{y}$ (indeks performa siswa),
4) kenaikan satu standar deviasi pada x_4 (waktu tidur siswa) menyebabkan peningkatan sebesar 0.029 standar deviasi pada $\hat{y}$ (indeks performa siswa),
5) kenaikan satu standar deviasi pada x_5 (banyaknya contoh soal yang dikerjakan) menyebabkan peningkatan sebesar 0.33 standar deviasi pada $\hat{y}$ (indeks performa siswa),
6) peubah x_2 (nilai sebelumnya) memiliki pengaruh yang paling besar terhadap peubah $\hat{y}$ (indeks performa siswa) karena nilai koefisiennya paling besar, dan
7) peubah x_4 (waktu tidur siswa) memiliki pengaruh yang paling kecil terhadap peubah $\hat{y}$ (indeks performa siswa) karena nilai koefisiennya paling kecil.

Evaluasi Model Regresi

Model regresi kuadrat-terkecil yang diperoleh sebelumnya dievaluasi menggunakan data latih dan data uji. Terdapat tiga kriteria yang digunakan di dalam evaluasi model, yaitu *R-squared* untuk data uji, *Root Mean Squared Error* (RMSE) untuk data latih dan data uji, dan *Mean Absolute Error* (MAE) untuk data latih dan data uji. Berikut ini disajikan nilai masing-masing kriteria evaluasi model pada Tabel 5.

Tabel 5. Hasil Evaluasi Model Regresi Kuadrat-Terkecil

Kriteria	Nilai
R^2 Data Uji	0.988
RMSE Data Latih	0.106
RMSE Data Uji	0.108
MAE Data Latih	0.084
MAE Data Uji	0.086

Berdasarkan Tabel 5, nilai R^2 yang diperoleh mendekati 1, artinya model yang diperoleh dapat menjelaskan sebagian besar variasi dalam peubah respon (indeks performa siswa) pada data uji. Kemudian, jika dilihat dari nilai RMSE untuk data latih dan data uji, model tersebut dapat memprediksi data latih dan data uji dengan baik. Selanjutnya, jika dilihat dari nilai MAE untuk data latih dan data uji, model tersebut memiliki galat rata-rata yang rendah pada data latih maupun data uji. Selain itu, nilai RMSE untuk data latih serupa dengan nilai RMSE untuk data uji, dan nilai MAE untuk latih juga serupa dengan nilai MAE untuk data uji, artinya model tidak mengalami *overfitting*. Nilai RMSE dan MAE pada kedua data cukup rendah, sehingga model memiliki kemampuan generalisasi yang baik.

PENUTUP

Simpulan

Penelitian ini menggunakan teknik *machine learning* dalam membangun model regresi kuadrat-terkecil untuk memprediksi indeks performa siswa berdasarkan waktu belajar, nilai sebelumnya, keikutsertaan dalam kegiatan ekstrakurikuler, waktu tidur, dan banyaknya contoh soal yang dikerjakan. Berdasarkan model yang diperoleh, faktor yang paling berpengaruh terhadap indeks performa siswa adalah nilai sebelumnya, sedangkan faktor yang paling tidak berpengaruh adalah waktu tidur siswa. Berdasarkan kriteria evaluasi model, model dapat memprediksi data dengan baik, memiliki galat rata-rata yang rendah pada data yang digunakan, tidak mengalami *overfitting*, dan memiliki kemampuan generalisasi yang baik.

Saran

Penelitian ini membahas tentang salah satu model *machine learning*, yaitu regresi kuadrat-terkecil yang diterapkan untuk melakukan prediksi indeks performa siswa berdasarkan beberapa faktor yang diduga memiliki pengaruh. Untuk penelitian selanjutnya, dapat digunakan model *machine learning* yang lain untuk menyelesaikan permasalahan yang sama, seperti *ridge regression*, *huber regression*, *quantile regression*,

dan lain-lain. Selain itu, faktor-faktor lainnya selain yang digunakan pada penelitian ini juga dapat ditambahkan sebagai peubah prediktor.

DAFTAR PUSTAKA

Altabrawee, H., Ali, O. A. J., & Ajmi S. Q. (2019). Predicting students' performance using machine learning techniques. *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences*, 27(1), 194-205.

Angra, S., & Ahuja, S. (2017). Machine learning and its applications: A review. *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, 57-60.

Asthana, P., Mishra, S., Gupta, N., Derawi, M., & Kumar, A. (2023). Prediction of student's performance with learning coefficients using regression based machine learning models. *IEEE Access*, 11, 72732-72742.

Gu, M., Liu, Y., Sun, H., Sun, H., Fang, Y., Chen, L., & Zhang, L. (2024). Using machine learning to predict the risk of short-term and long-term death in acute kidney injury patients after commencing CRRT. *BMC Nephrology*. 25(1).

Haldorai, A., Ramu, A., & Suriya, M. (2020). *Organization Internet of Things (IoTs): Supervised, Unsupervised, and Reinforcement Learning*. Di dalam: Haldorai A, Ramu A, Khan S, editor. *Business Intelligence for Enterprise Internet of Things*. Springer Cham.

Htun, H. H., Biehl, M., & Petkov, N. (2024). Forecasting relative returns for S&P 500 stocks using machine learning. *Financial Innovation*. 10(1).

Huang, Z. C., Sangiorgi, I., & Urquhart, A. (2024). Forecasting Bitcoin volatility using machine learning techniques. *Journal of International Financial Markets, Institutions and Money*. 97.

Hussain, S., & Khan, M. Q. (2023). Student-performulator: predicting students' academic performance at secondary and intermediate level using machine learning. *Ann. Data. Sci*. 10, 637-655.

Jalal, D., & Ezzedine, T. (2019). Performance analysis of machine learning algorithms for water quality monitoring system. *2019 International Conference on Internet of Things, Embedded Systems and Communications, IINTEC 2019 - Proceedings*, 86–89.

Mamo, D. N., Gebremariam, Y. H., Adem, J. B., Kebede, S. D., & Walle, A. D. (2024). Machine learning to predict unintended pregnancy among reproductive-age women in Ethiopia: evidence from EDHS 2016. *BMC Women's Health*. 24(1).

Morales, E. F., & Escalante, H. J. (2022). *Chapter 6 - A brief introduction to supervised, unsupervised, and reinforcement learning*. Di dalam: Torres-García AA, Reyes-García CA, Villaseñor-Pineda L, Mendoza-Montoya O, editor. *Biosignal Processing and Classification Using Computational Learning and Intelligence*. London: Academic Press. hlm 111-129.

Nixon, P., & Gilbert, E. (2024). Using machine learning to predict investors' switching behaviour. *Journal of Behavioral and Experimental Finance*. 44.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*. 80(3), 3782-3785.

Sutton, R. S. (1992). Introduction: The challenge of reinforcement learning. Machine Learning. 8(3-4), 22-227.

Valkenborg, D., Rousseau, A. J., Geubbelmans, M., & Burzykowski, T. (2023). Unsupervised learning. *Am J Orthod Dentofacial Orthop*. 163(6), 877-882.

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*. 9(11).

Zhang, Y., Ma, H., Wang, H., Xia, Q., Wu, S., Meng, J., Zhu, P., Guo, Z., & Hou, J. (2024). Forecasting the trend of tuberculosis incidence in Anhui Province based on machine learning optimization algorithm, 2013–2023. *BMC Pulmonary Medicine*. 24(1).