

STUDENT SCORE PREDICTION

Muhammad Naufal FT





LATAR BELAKANG

Regresi adalah salah satu metode dalam pembelajaran mesin (machine learning) yang digunakan untuk memodelkan hubungan antara variabel independen dan dependen. Dalam konteks ini, Anda mungkin sedang bekerja dengan regresi linier atau random forest regression, yang sering digunakan untuk prediksi numerik.

Proyek mini portofolio ini dirancang untuk menelusuri dan memprediksi kinerja siswa berdasarkan pola belajar mereka.

Secara lebih spesifik, proyek ini memiliki tujuan-tujuan berikut:

- 01.** Menganalisis hubungan antara jumlah jam belajar dan nilai yang diperoleh oleh siswa dalam ujian.
- 02.** Membandingkan berbagai model regresi (Regresi Linier dan Decision Tree Regressor) serta mengevaluasi kinerjanya.
- 03.** Melakukan analisis data melalui statistik ringkasan dan visualisasi untuk memperoleh wawasan dari dataset.
- 04.** Menentukan model yang paling akurat untuk memprediksi nilai siswa, serta menarik kesimpulan mengenai efektivitas model pada dataset berukuran kecil.
- 05.** Menerapkan model regresi machine learning untuk memprediksi nilai berdasarkan jam belajar menggunakan data nyata.

TUJUAN PROYEK



EXPLORATORY DATA ANALYSIS (EDA)

Sebelum memasuki tahap pemodelan, penting untuk memahami isi dataset terlebih dahulu, melalui proses Exploratory Data Analysis (EDA), kita dapat menggali struktur data, membantu kita mengenali karakteristik data dan memastikan data siap digunakan dalam model.

Dataset = Student Scores

Isi Dataset:

1. **hours** = total jumlah jam belajar
2. **scores** = total nilai ujian

Descriptive Statistics:

	hours	scores
count	25.000000	25.000000
mean	5.012000	51.480000
std	2.525094	25.286887
min	1.100000	17.000000
25%	2.700000	30.000000
50%	4.800000	47.000000
75%	7.400000	75.000000
max	9.200000	95.000000

Correlation:

	hours	scores
hours	1.000000	0.976191
scores	0.976191	1.000000



CHECK DUPLICATED DATA

```
Jumlah data duplikat: 0
```

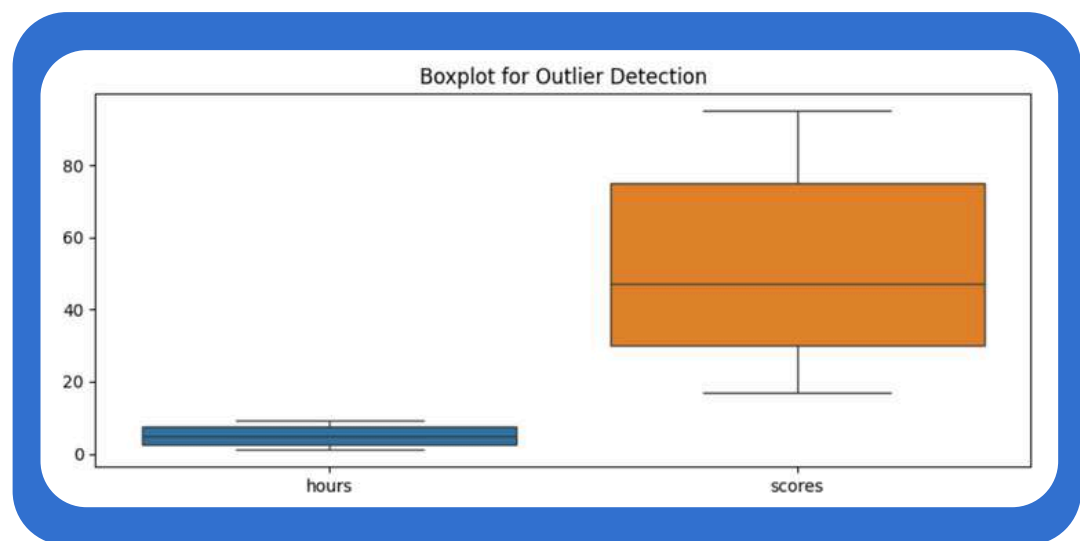
Tidak ada data set yang diduplikat

CHECK MISSING VALUE HANDLING

```
Cek missing value:  
hours      0  
scores     0  
dtype: int64
```

Tidak ada missing value

OUTLIER ANALYSIS



outlier tidak ditemukan

TRAINING DAN EVALUASI MODEL

Linear Regression Evaluation:

MSE: 18.943211722315272

RMSE: 4.352380006653288

R2 Score: 0.9678055545167994

Decision Tree Regressor Evaluation:

MSE: 31.7

RMSE: 5.630275304103699

R2 Score: 0.9461250849762066

Model yang digunakan:

- linear Regression
- Decision Tree

Evaluasi dengan:

- Mean squared error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared (R2)



HASIL EVOLUSI MODEL DARI DATA STUDENT SCORE

Linear Regression MSE: 18.94 | RMSE: 4.35 | R2: 0.97

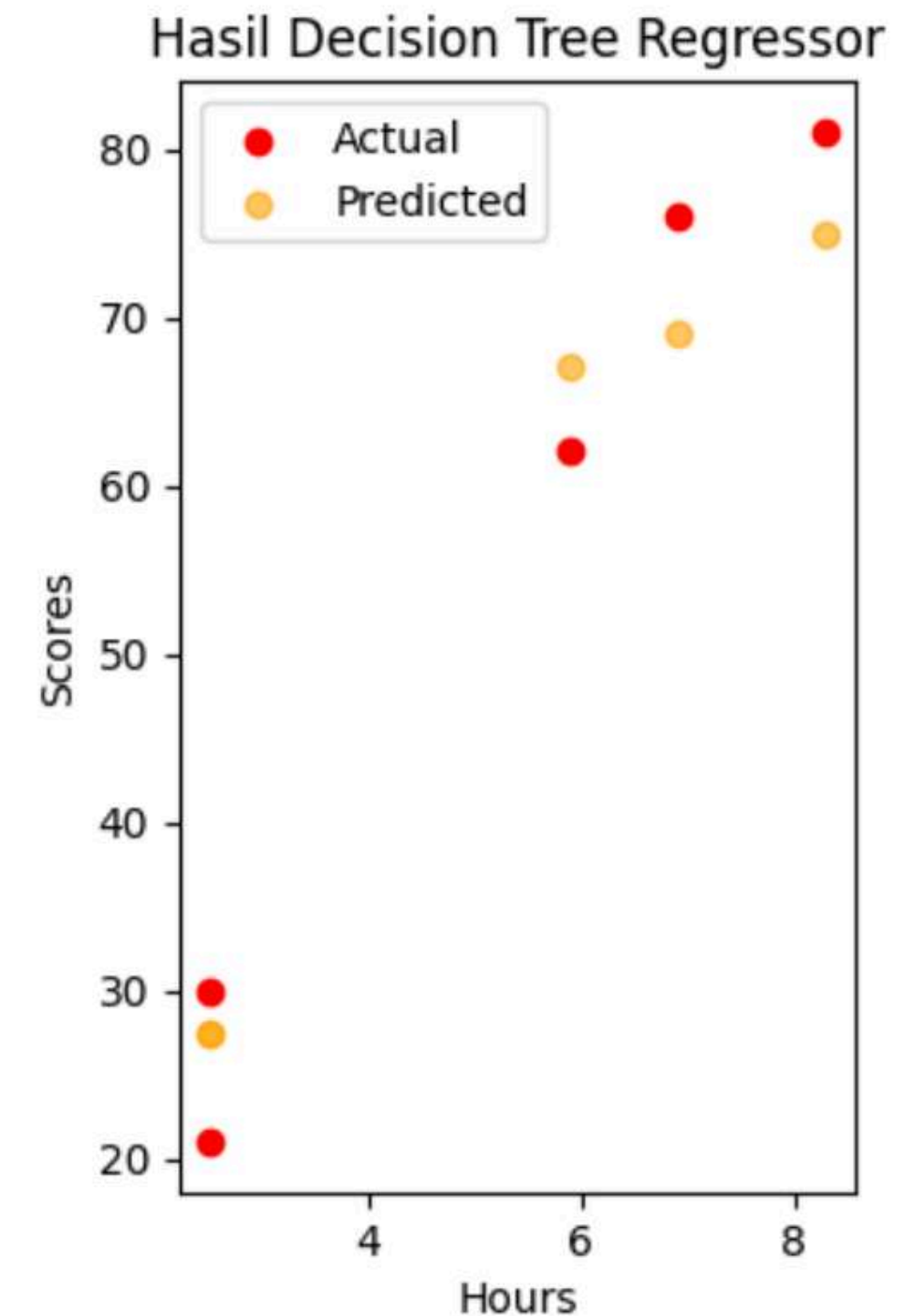
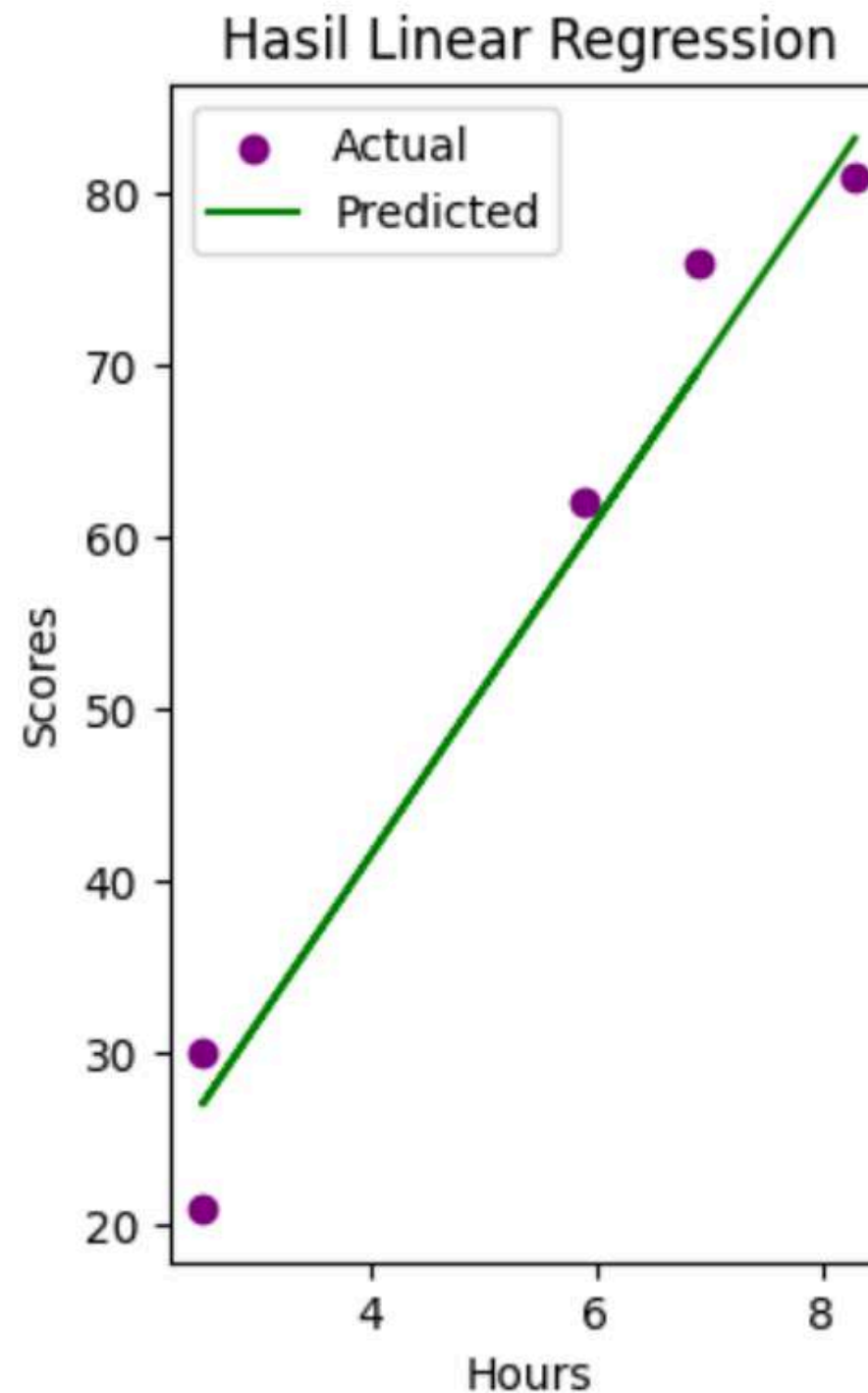
Decision Tree MSE: 31.70 | RMSE: 5.63 | R2: 0.95

Berdasarkan hasil analisis, model Linear Regression direkomendasikan sebagai metode prediksi nilai ujian dalam kasus ini. Model ini tidak hanya memberikan hasil yang akurat, tetapi juga mudah diterapkan serta sangat cocok dengan karakteristik data yang digunakan.



MODEL COMPARASION

Berikut hasil model comparison untuk **Model Linear Regression** dan **Model Decision Tree Regressor**



KESIMPULAN



Setelah dilakukan analisis data dan pembangunan model regresi, berikut adalah beberapa kesimpulan yang dapat diambil:

- Terdapat hubungan linear yang sangat kuat antara jumlah jam belajar dan nilai ujian, dengan nilai korelasi sekitar 0,98.
- Model Linear Regression menunjukkan kinerja yang lebih baik dibandingkan Decision Tree Regressor berdasarkan semua metrik evaluasi (MSE, RMSE, dan R^2).
- Model regresi linear sangat sesuai untuk digunakan pada dataset yang sederhana dan memiliki pola hubungan linear yang kuat.
- Dataset yang digunakan tergolong bersih, tidak terdapat data yang hilang maupun outlier yang signifikan.
- Dua model telah diterapkan dalam analisis ini, yaitu Linear Regression dan Decision Tree Regressor.



THANK
YOU

