**Nama :** Arvin Azmi Sava

**No. Peserta : 111**

**PROJECT ANALISIS DATA DAC 2023**

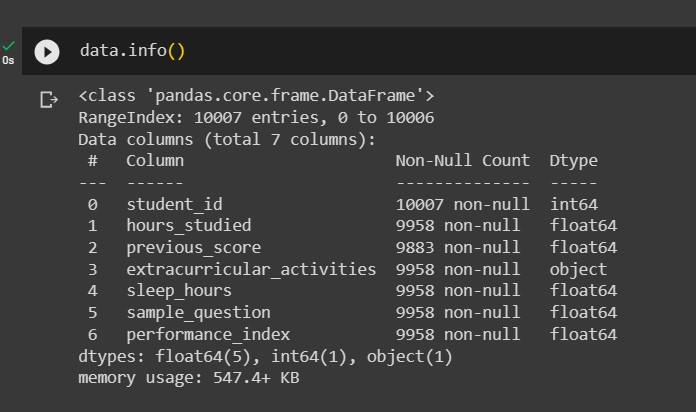
**Tujuan dan Batasan Masalah Analisis**

Tujuan dari analisis ini adalah untuk menjawab beberapa pertanyaan berikut:

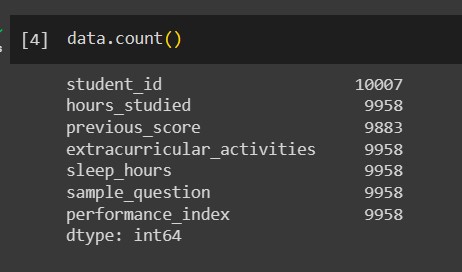
1. Bagaimana pengaruh lama waktu belajar siswa terhadap indeks performa siswa?
2. Bagaimana pengaruh jam tidur siswa terhadap indeks performa siswa?

Adapun batasan masalah analisis ini adalah hanya menganalisis variabel lama waktu belajar dan lama waktu jam tidur terhadap performa indeks siswa.

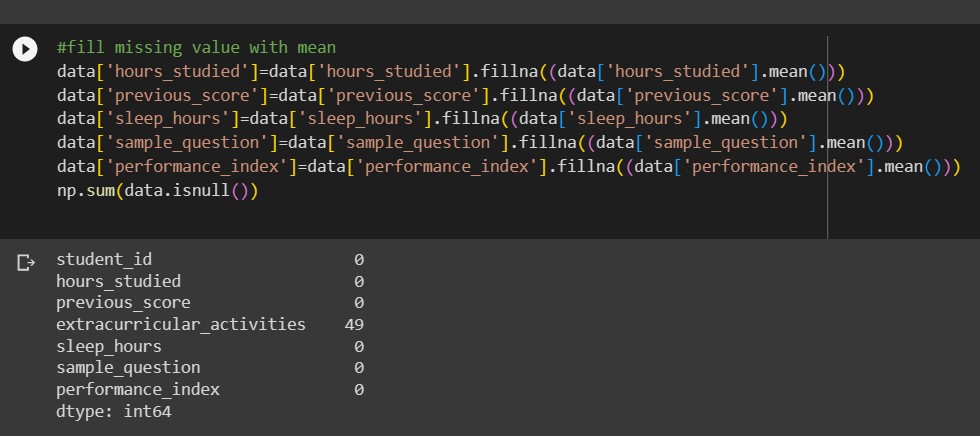
**Pengetahuan Data**

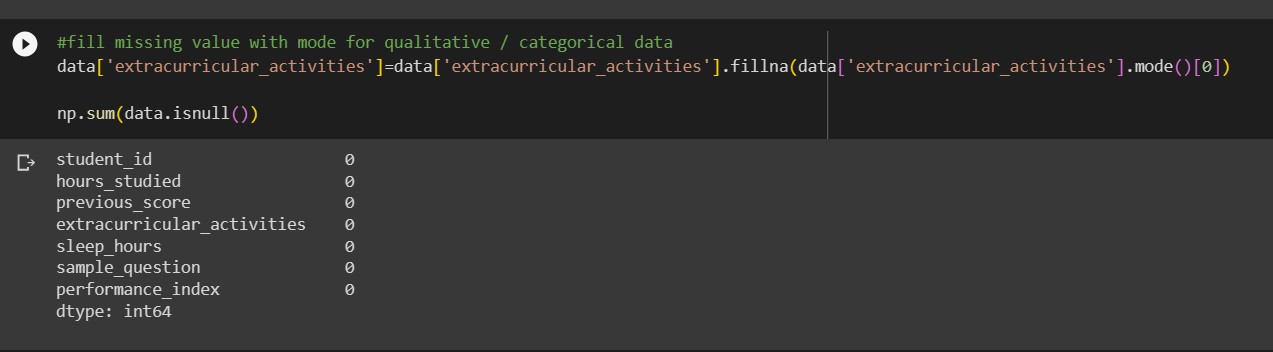
Dalam datasets, terdapat 7 kolom yaitu student\_id, hours\_studied, previous\_score, extracurricular\_activities, sleep\_hours, sample\_question, performance\_index. Untuk tipe data masing-masing kolom ditunjukkan pada gambar di bawah ini

**Persiapan Data**

Pada tahap ini diperiksa apakah dataset tersebut terdapat missing value atau tidak. Setelah diperiksa, ternyata ditemukan dari dataset tersebut terdapat missing value yaitu terjadi ketidaksesuain jumlah antara masing-masing kolom. 

Oleh karena itu perlu kita lakukan teknik imputasi yaitu mengganti missing value tersebut dengan menggunakan median atau mean. Untuk menentukan menggunakan median atau mean, harus diperiksa terlebih dahulu apakah terdapat outlier atau tidak. Jika ada outlier, maka menggunakan median. Jika tidak ada, maka menggunakan mean. Setelah dilakukan pemeriksaan, ternyata tidak ditemukan outlier sehingga kita dapat mengganti missing value dengan menggunakan nilai mean

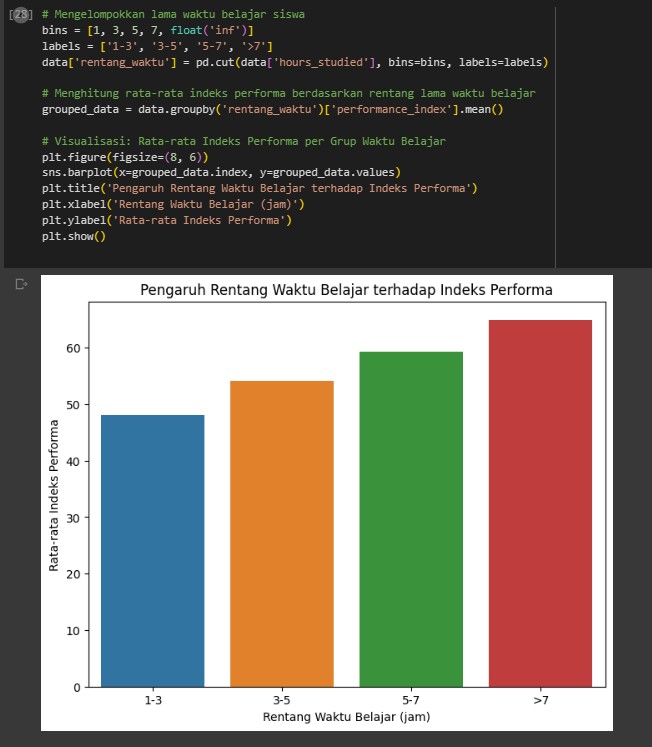


Untuk kolom extracurricular\_activites karena berupa tipe data objek atau yang menunjukkan kategori, maka missing valuenya akan diganti dengan nilai yang sering muncul (modus).

Dataset tersebut kini sudah bersih dan bisa dilakukan proses analisis eksplorasi

**Analisis Eksplorasi Data**

1. **Pengaruh waktu lama belajar terhadap indeks performa**

Untuk mengetahui pengaruh waktu lama belajar terhadap indeks performa, Hal pertama yang saya lakukan adalah membagi lama waktu belajar siswa menjadi beberapa interval yaitu 1-3 jam, 3-5 jam, 5-7 jam, dan lebih dari 7 jam. kemudian setelah itu, saya menghitung rata-rta Indeks Performa(IP) dari masing-masing kelompok interval tersebut dengan menggunakan atribut groupby. Kemudian, hasilnya saya visualisasikan menggunakan diagram batang / barplot seperti gambar di bawah ini. Secara sekilas dapat terlihat bahwa siswa yang jam belajarnya lebih lama mendapatkan indeks performa yang lebih tinggi. Namun, untuk mengetahui seberapa kuat penaruh tersebut akan dijelaskan pada langkah selanjutnya

1. **Pengaruh lama jam tidur terhadap indeks performa siswa**

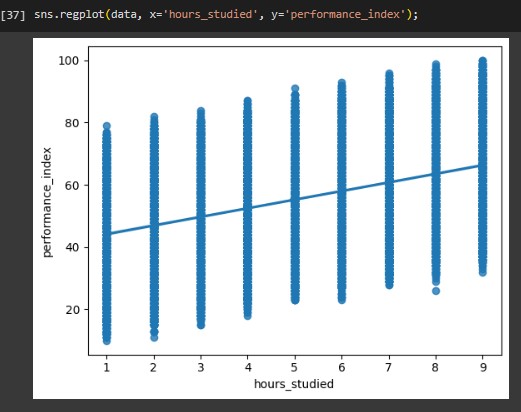
Di sini saya menggunakan cara yang sama seperti sebelumnya, hanya saja variabel lama waktu belajar diganti dengan lama waktu tidur. Kemudian hasilnya divisualisasikan menggunakan barplot seperti gambar di bawah ini. Dapat dilihat bahwa lama jam tidur siswa tidak berpengaruh signifikan terhadap indeks performa siswa.

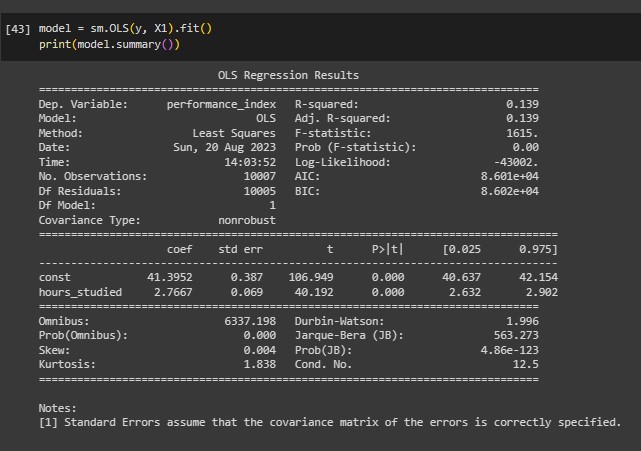


**Modelling**

1. **Pengaruh waktu lama belajar terhadap indeks performa**

Untuk mengetahui seberapa pengaruh lama waktu belajar terhadap indek performa, maka saya visualisasikan hubungan kedua variabel tersebut dan didapatkan hasil seperti gambar di bawah ini

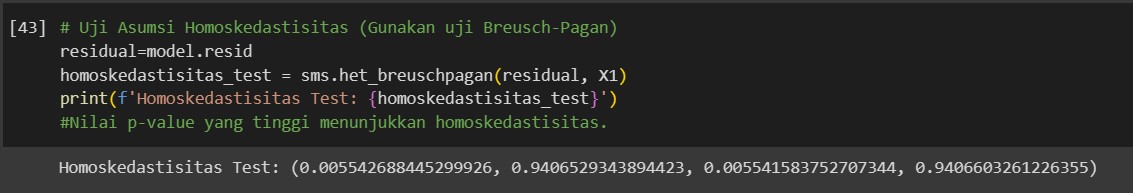


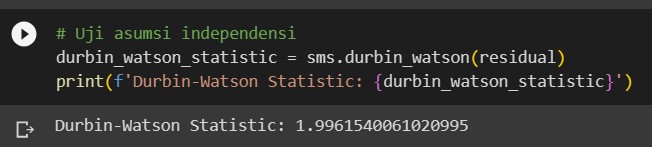
Setelah itu, dilakukan analisis regresi untuk menilai seberapa berpengaruh hubungan variabel tersebut dan didapatkan hasil seperti di bawah ini

R-squared sebesar 0.139 menunjukkan bahwa variabel "hours\_studied" berpengaruh sebesar 13.9% terhadap variabel "performance\_index".

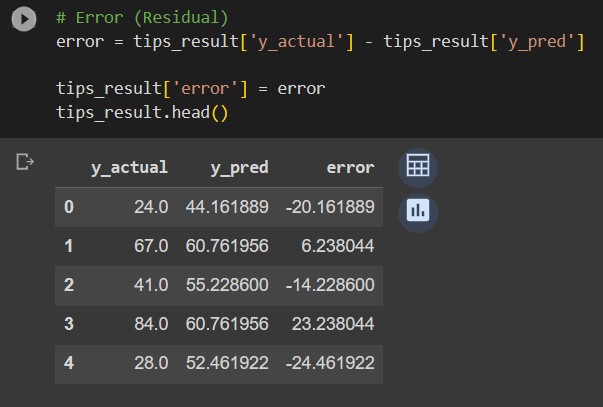
**Model Evaluation**

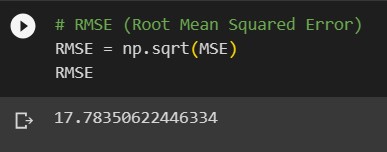
1. **Pengaruh waktu lama belajar terhadap indeks performa**

Di sini saya menggunakan beberapa uji untuk mengevaluasi model. Pertama, saya gunakan Uji Asumsi Homoskedastisitas (Uji Breusch-Pagan) untuk menguji bahwa varians dari residual (kesalahan) model regresi konstan dan tidak bergantung pada nilai-nilai prediktor. Hasil dari uji ini sebagai berikut. Nilai p-value di sini dapat dikatakan sangat rendah menunjukkan bahwa asumsi homoskedastisitas belum terpnuhi

Selanjutnya, saya gunakan Uji Asumsi Independensi (Durbin-Watson Statistik) yang menghasilkan nilai sekitar 2. Hal ini menunjukkan bahwa asumsi independensi cukup terpenuhi

Kemudian, langkah berikutnya dilakukan untuk memprediksi nilai. Hasil dari prediksi nilai ini akan dibandingkan dengan nilai aktual. Berikut ini adalah perbandingan nilai prediksi dan nilai aktual



Langkah Terkahir yaitu melakukan Perhitungan Mean Squared Error (MSE) dan Root Mean Squared Error (RMSE). MSE dihitung sebagai rata-rata dari kesalahan kuadrat, sedangkan RMSE dihitung sebagai akar kuadrat dari MSE. MSE dan RMSE digunakan untuk mengukur seberapa baik model prediksi bekerja dalam mengestimasi nilai aktual. Berikut ini adalah hasil perhitungan Root Mean Squared Error (RMSE)

Terlihat bahwa Nilai RMSE cukup tinggi yang menunjukkan bahwa model ini memiliki kinerja yang kurang baik.

**Kesimpulan dan Saran**

Berdasarkan proses analisis data yang dilakukan dari awal sampai akhir, dapat disimpulkan:

1. Variabel lama waktu belajar siswa mempengaruhi indeks performa akhir siswa. Hal ini ditunjukkan dengan korelasi kedua variabel tersebut yang cukup tinggi yaitu variabel “hours\_studied” mempengaruhi sebesar 37,28% variabell “performance\_index” sehingga dapat dikatakan siswa yang belajar lebih lama kemungkinan akan mendapatkan indek performa akhir yang lebih tinggi.
2. Variabel lama waktu tidur siswa tidak berpengaruh signifikan terhadap indeks performasi akhir siswa. Hal ini ditunjukkan dengan korelasi kedua variabel tersebut yang cukup rendah yaitu variabel “sleep\_hours” mempengaruhi hanya sebesar 4,79% variabel “performance\_index” sehingga dapat dikatakan bahwa lama waktu tidur siswa tidak berpengaruh signifikan terhadap indeks performa.

Saran : Dari kesimupulan tersebut sebaiknya siswa harus belajar lebih lama agar mendapatkan nilai akhir yang memuaskan

**Daftar Pustaka (Jika Ada)**

**Lampiran (Jika Ada)**

Disclaimer: Saya sejujurnya kurang memahami materi Day 4 bagian pembuatan model dan evaluasi model karena background saya berasal dari jurusan Sistem Informasi. Semua syntax yang saya tulis pada bagian tersebut meniru punya Mas Unggul sehingga saya kurang tahu hal apa yang harus dilakukan apabila saat dilakukan uji asumsi menunjukkan pembuatan model masih belum cukup terpenuhi dan menunjukkan error yang masih tinggi.