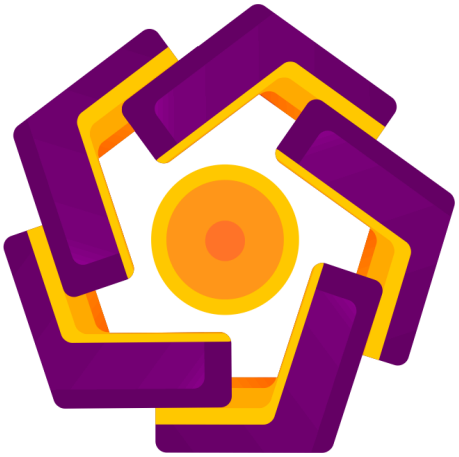
Big data & Predictive Analytics Final project

Prediksi GPA Berdasarkan Study Time Weekly, Absences, Tutoring, dan Grade Class

Dosen pengampu:
Mulia Sulistiyono, S.Kom., M.Kom



Anggota kelompok:

1. Vianda Retnaningtiyas Purbandari Karetji, 23.11.5445
2. Rifky Danu Asmoro, 23.11.5489
3. Stefanus Arya Bayu Samudra Bataona, 23.11.5477

Program Studi Informatika
Fakultas Ilmu Komputer
Universitas Amikom Yogyakarta

2025

Daftar Isi

1. Latar belakang

Pendidikan merupakan salah satu fondasi utama dalam pembangunan sumber daya manusia. Dalam dunia pendidikan, prestasi akademik siswa menjadi tolak ukur keberhasilan proses belajar mengajar. Salah satu indikator yang paling umum digunakan untuk mengukur prestasi tersebut adalah Grade Point Average (GPA) atau Indeks Prestasi Kumulatif. GPA mencerminkan hasil belajar siswa secara menyeluruh dalam jangka waktu tertentu, dan menjadi pertimbangan penting dalam banyak aspek, seperti kelulusan, penerimaan beasiswa, hingga seleksi masuk perguruan tinggi. Oleh karena itu, penting untuk memahami faktor-faktor yang mempengaruhi GPA agar dapat meningkatkan kualitas pendidikan secara menyeluruh.

Banyak penelitian menunjukkan bahwa prestasi akademik dipengaruhi oleh berbagai faktor, mulai dari faktor internal seperti motivasi belajar, kemampuan intelektual, hingga faktor eksternal seperti lingkungan sosial, ekonomi keluarga, dan ketersediaan fasilitas pendidikan. Salah satu faktor yang cukup dominan dan dapat diukur secara objektif adalah jumlah waktu belajar siswa setiap minggu. Waktu belajar yang cukup secara konsisten diyakini dapat meningkatkan pemahaman terhadap materi pelajaran, yang pada akhirnya berdampak pada kenaikan nilai akademik. Namun demikian, belum tentu setiap siswa yang belajar dalam waktu lama otomatis memperoleh nilai tinggi. Oleh karena itu, perlu dilakukan analisis statistik untuk mengetahui sejauh mana waktu belajar mempengaruhi prestasi akademik siswa.

Melalui proyek ini, dilakukan analisis dan prediksi GPA berdasarkan jumlah jam belajar per minggu, jumlah ketidakhadiran siswa, bimbingan dan grade class menggunakan pendekatan regresi linier sederhana. Tujuan dari penelitian ini adalah untuk mengetahui seberapa besar pengaruh waktu belajar terhadap GPA, serta membuat model prediksi yang dapat memperkirakan GPA siswa berdasarkan input waktu belajar, jumlah ketidakhadiran siswa, bimbingan, dan grade class. Dengan adanya model ini, guru, orang tua, dan siswa dapat memiliki acuan yang berbasis data dalam menyusun strategi belajar yang lebih efektif. Selain itu, proyek ini juga memberikan gambaran nyata tentang penerapan data science dalam bidang pendidikan, serta menjadi contoh konkret bagaimana data dapat digunakan untuk pengambilan keputusan yang lebih baik.

2. Metode

A. Exploratory Data Analysis (EDA)

Digunakan untuk memahami dan memvisualisasikan data.
Metode/teknik yang dipakai:

- Histogram **(**sns.histplot**)** untuk melihat distribusi nilai GPA.
- Boxplot (sns.boxplot) untuk melihat distribusi GPA berdasarkan kategori Gender.
- Heatmap (sns.heatmap) untuk melihat korelasi antar variabel numerik di dataset.
- Scatterplot (sns.scatterplot) untuk melihat hubungan antara Study Time dan GPA.
- Countplot (sns.countplot) untuk melihat jumlah siswa berdasarkan kategori Ethnicity.

B. Simple & Multiple Linear Regression (Model Prediksi GPA)

- Multiple Linear Regression menggunakan LinearRegression() dari sklearn.
  Rumus modelnya seperti ini:
  *GPA= intercept + (coef_studytime * studytime) + (coef_absence * absences) + (coef_tutoring * tutoring) + (coef_gradeclass * gradeclass)*.
- Training & Testing Split:
  X_train, X_test, y_train, y_test = train_test_split(...)
  Data dibagi 80% training dan 20% testing.
- Model Fitting : *model.fit(X_train, y_train)*
- Evaluasi Model :
  a. Root Mean Squared Error (RMSE) ➔ Mengukur seberapa jauh prediksi dari nilai aktual.

$$RMSE = \sqrt{\frac{1}{n}\sum(y_{pred} - y_{actual})^2}$$

  b. R^2 Score ➔ Mengukur seberapa baik model menjelaskan variasi data (fit model).

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

  c. Prediksi Manual di Collab / predict : User mengisi input studytime dan absences
  Sistem menghitung GPA dengan rumus linear regression manual dari hasil model :
  *gpa = intercept + (coef_studytime * studytime) + (coef_absence * absences) + (coef_tutoring * tutoring) + (coef_gradeclass * gradeclass)*

2.1 Alur final project

A. Data Collection (Pengumpulan Data)

Tahapan pertama dimulai dengan mengimpor dataset dari file CSV bernama Student_performance_data.csv. Dataset ini berisi informasi mengenai data pribadi dan akademik siswa, seperti:

a. Usia (Age)
b. Gender (Gender)
c. Jam belajar per minggu (StudyTimeWeekly)
d. Kehadiran (Attendance)
e. Nilai akhir (GPA)
f. Kategori kelas nilai (GradeClass)

Data dibaca menggunakan pustaka pandas melalui perintah pd.read_csv(). Setelah data dimuat, dilakukan pemeriksaan awal seperti df.head(), df.info(), dan df.describe() untuk melihat struktur dan deskripsi statistik dasar dataset.

B. Exploratory Data Analysis (EDA) dan Visualisasi Data

Tahapan ini bertujuan untuk memahami karakteristik data lebih dalam dan mendeteksi pola awal. Beberapa proses yang dilakukan antara lain:

a. Mengecek Missing Value: Menggunakan df.isnull().sum() untuk memastikan tidak ada nilai kosong yang signifikan.
b. Distribusi GPA: Visualisasi menggunakan histogram dan KDE (sns.histplot) untuk melihat pola nilai akhir siswa.
c. Boxplot GPA berdasarkan Gender: Untuk membandingkan distribusi nilai antara siswa laki-laki dan perempuan menggunakan sns.boxplot.
d. Distribusi Grade Class: Dengan sns.countplot, dilihat sebaran siswa berdasarkan kelas nilai yang mereka peroleh.
e. Scatter Plot: Visualisasi hubungan antara StudyTimeWeekly dan GPA dengan sns.scatterplot, yang memperlihatkan tren awal hubungan linier.

Tahapan EDA ini sangat penting sebagai dasar pemodelan dan untuk mengidentifikasi outlier atau pola yang tidak biasa.

C. Analisis Korelasi

Pada tahap ini, dilakukan analisis korelasi antar variabel numerik menggunakan heatmap korelasi dari pustaka seaborn. Fungsi df.corr() digunakan untuk menghitung nilai korelasi Pearson antar fitur seperti:

a. StudyTimeWeekly
b. Attendance
c. GPA

Visualisasi dilakukan menggunakan sns.heatmap() dengan anotasi dan skema warna yang memperjelas kekuatan hubungan. Hasilnya menunjukkan bahwa terdapat korelasi positif antara waktu belajar dan GPA, yang menjadi dasar untuk membangun model prediksi.

D. Membuat Model Regresi Linier

Berdasarkan hasil korelasi, dibangun model Regresi Linier Sederhana untuk memprediksi GPA berdasarkan StudyTimeWeekly.

a. Data dibagi menjadi data latih dan data uji menggunakan train_test_split() dari pustaka scikit-learn.
b. Model dilatih menggunakan LinearRegression().fit() dari sklearn.linear_model.
c. Setelah dilatih, diperoleh nilai:

- Intercept (β_0): Nilai GPA saat jam belajar = 0
- Koefisien (β_1): Menunjukkan pengaruh tambahan GPA setiap penambahan 1 jam belajar

Model ini membentuk sebuah garis regresi yang mendekati pola sebaran data pada scatter plot.

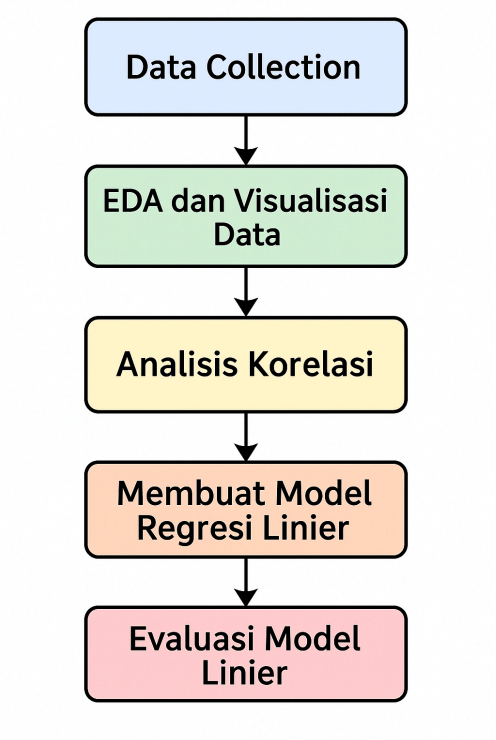
E. Evaluasi Model Linier

Setelah model dibuat, langkah selanjutnya adalah mengevaluasi performa prediksi model menggunakan metrik Root Mean Squared Error (RMSE).

a. Nilai RMSE dihitung menggunakan fungsi mean_squared_error() dari sklearn.metrics.
b. Semakin kecil nilai RMSE, semakin baik performa model dalam memprediksi GPA.
c. Visualisasi dilakukan dengan:

- Scatter plot data uji (X_test, y_test) sebagai data asli.
- Garis regresi dari y_pred untuk membandingkan hasil prediksi dengan data sebenarnya.

- Flowchart


Data Collection
EDA dan Visualisasi Data
Analisis Korelasi
Membuat Model Regresi Linier
Evaluasi Model Linier

2.2 Dataset

Data yang digunakan dalam analisis ini diperoleh dari situs Kaggle, menggunakan dataset berjudul Students Performance Dataset (https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset/data) Dataset ini dipilih karena relevan dengan tujuan Final Project UAS, yaitu membangun model prediksi dan klasifikasi untuk memahami faktor-faktor yang memengaruhi prestasi akademik (GPA). Dataset tersebut mencakup berbagai informasi yaitu :

a. StudentsID : kode unik yang diberikan kepada masing-masing siswa untuk membedakan satu sama lain dalam sistem data.
b. Age : usia para siswa berada dalam rentang antara 15 hingga 18 tahun.
c. Gender : jenis kelamin siswa dimana 0 mewakili laki-laki dan 1 mewakili perempuan
d. Ethnicity : latar belakang etnis. Dibagi menjadi 4 yaitu :
   0 : merepresentasikan siswa dengan latar belakang etnis Caucasian.
   1 : merepresentasikan siswa dengan latar belakang etnis African American.
   2 : merepresentasikan siswa dengan latar belakang etnis Asian.
   3 : merepresentasikan siswa dengan latar belakang etnis Lainnya.
e. Parental Education : tingkat pendidikan yang ditempuh oleh orang tua siswa. Dibagi menjadi 5 yaitu :
   0 : tingkat pendidikan None (tidak disebutkan)
   1 : tingkat pendidikan High School (SMA)
   2 : tingkat pendidikan Some Collage (telah mengikuti pendidikan tinggi setelah SMA, tetapi belum menyelesaikan gelar sarjana)
   3 : tingkat pendidikan Bachelors (S1)
   4 : tingkat pendidikan Higher (Tingkatan diatas S1)
f. StudyTimeWeekly : waktu belajar siswa setiap minggu, dihitung dalam jam.
g. Absences : Jumlah ketidakhadiran siswa selama tahun ajaran.
h. Tutoring : Status les atau bimbingan belajar, di mana angka 0 berarti Tidak mengikuti les, dan 1 berarti Mengikuti les.
i. Parental Support : menggambarkan tingkat dukungan emosional, akademik, atau finansial yang diberikan orang tua kepada anaknya. Dibagi menjadi 5 yaitu :
   0 : None (tidak ada parental support)
   1 : Low (tingkat parental support rendah)
   2 : Moderate (tingkat parental support rendah)
   3 : High (tingkat parental support tinggi)
   4 : Very High (tingkat parental support sangat tinggi)
j. Extracurricular : partisipasi dalam kegiatan ekstrakurikuler, di mana 0 berarti Tidak dan 1 berarti Ya
k. Sports : partisipasi dalam olahraga, di mana 0 berarti Tidak dan 1 berarti Ya
l. Music : partisipasi dalam kegiatan musik, di mana 0 berarti Tidak dan 1 berarti Ya
m. Volunteering : Partisipasi dalam kegiatan sukarela, di mana 0 berarti Tidak dan 1 berarti Ya
n. GPA : Grade Point Average (nilai rata-rata prestasi akademik) berada dalam rentang 2.0 hingga 4.0, dan dipengaruhi oleh kebiasaan belajar

2.3 EDA

1. **Pemeriksaan Struktur dan Statistik Data**

   Langkah pertama dalam tahap EDA adalah memahami struktur data secara keseluruhan. Hal ini dilakukan dengan menampilkan beberapa baris awal dari dataset menggunakan fungsi df.head() untuk melihat isi dan format data. Selanjutnya, digunakan df.info() untuk mengetahui tipe data di setiap kolom, jumlah data non-null, dan struktur tabel. Dengan menggunakan df.describe(), diperoleh ringkasan statistik seperti nilai minimum, maksimum, rata-rata, dan kuartil pada kolom numerik seperti GPA, Age, dan StudyTimeWeekly. Informasi ini penting untuk mengetahui sebaran awal dan mendeteksi kemungkinan anomali seperti nilai ekstrim atau ketidakwajaran dalam data.

2. **Pengecekan Missing Value**

   Salah satu langkah penting dalam pembersihan data adalah memastikan bahwa tidak terdapat nilai yang hilang (missing values) yang dapat mengganggu proses analisis atau pelatihan model. Untuk itu, digunakan fungsi df.isnull().sum() yang akan menghitung jumlah data kosong di setiap kolom. Pemeriksaan lanjutan juga dilakukan dengan df[df.isnull().any(axis=1)] untuk melihat baris mana saja yang memiliki nilai kosong. Berdasarkan hasil analisis, dataset tidak mengandung nilai kosong yang signifikan, sehingga proses imputasi atau penghapusan data tidak diperlukan.

3. **Visualisasi Distribusi GPA**

   Setelah data dinyatakan bersih, analisis dilanjutkan dengan visualisasi distribusi nilai GPA siswa. Visualisasi dilakukan menggunakan histogram dan kurva KDE melalui fungsi sns.histplot(). Tujuan dari visualisasi ini adalah untuk melihat persebaran GPA di seluruh siswa, apakah simetris, normal, atau terdapat kemiringan (skewness). Dari grafik tersebut, terlihat bahwa sebagian besar siswa memiliki GPA dalam rentang menengah hingga tinggi, dengan sebaran yang relatif normal. Hal ini memberikan indikasi awal bahwa model regresi linier mungkin dapat bekerja dengan baik pada data ini.

4. **Boxplot GPA Berdasarkan Gender**

   Untuk melihat apakah terdapat perbedaan nilai akademik berdasarkan jenis kelamin siswa, dilakukan visualisasi boxplot antara kolom Gender dan GPA menggunakan fungsi sns.boxplot(). Visualisasi ini menunjukkan distribusi, median, dan outlier dari GPA berdasarkan kategori gender (0 = laki-laki, 1 = perempuan). Hasilnya menunjukkan adanya sedikit perbedaan median GPA antara siswa laki-laki dan perempuan, meskipun secara keseluruhan distribusinya cukup mirip. Informasi ini membantu dalam memahami variasi nilai GPA berdasarkan demografi siswa.

5. **Visualisasi Korelasi antar Variabel**

   Untuk menganalisis hubungan antar variabel numerik dalam dataset, digunakan matriks korelasi dengan df.corr() dan divisualisasikan menggunakan sns.heatmap() dengan anotasi. Korelasi dihitung menggunakan metode Pearson. Dari hasil heatmap, diketahui bahwa terdapat korelasi positif yang cukup jelas antara StudyTimeWeekly dengan GPA, serta antara Attendance dengan GPA. Temuan ini menjadi landasan kuat bahwa waktu belajar dapat dijadikan variabel prediktor dalam pemodelan regresi.

6. **Hubungan Study Time dengan GPA (Scatter Plot)**

   Untuk memperkuat dugaan bahwa StudyTimeWeekly memiliki hubungan linier dengan GPA, dibuat visualisasi scatter plot menggunakan sns.scatterplot(). Scatter plot memperlihatkan bahwa terdapat kecenderungan pola naik antara jumlah jam belajar dan nilai GPA, yang mengindikasikan bahwa semakin banyak siswa belajar, maka semakin tinggi nilai GPA yang diperoleh. Pola ini sangat penting karena menjadi dasar pemilihan model regresi linier sederhana dalam tahap modeling.

7. **Distribusi Grade Class**

   Terakhir, dilakukan analisis terhadap distribusi kelas nilai siswa dengan menggunakan sns.countplot() pada kolom GradeClass. Visualisasi ini memberikan gambaran seberapa banyak siswa berada di kategori nilai tertentu (misalnya A, B, C, dll). Hal ini bermanfaat untuk memahami bagaimana distribusi kinerja akademik siswa di seluruh populasi dalam dataset.

3. Eksperimen

3.1 Proses Eksperimen

1. Mengimpor library untuk analisis data, visualisasi, pemodelan regresi linier, dan evaluasi model.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_squared_error, r2_score
```

2. Mengimpor Dataset dari Google Drive

```
from google.colab import drive
drive.mount('/content/drive')
dataset_path = '/content/drive/MyDrive/Dataset/Students/Student_performance_data.csv'
df = pd.read_csv(dataset_path)
df.head()
```

Menghubungkan Google Drive ke Google Colab, lalu membaca file dataset CSV dan menampilkannya dalam bentuk tabel (DataFrame) menggunakan pandas.

3. Menampilkan Informasi dan Statistik Data

```
print(df.info())
print(df.describe())
```

Menampilkan struktur data (df.info()) seperti jumlah kolom, tipe data, dan non-null entries, serta statistik deskriptif (df.describe()) seperti rata-rata, nilai minimum, maksimum, dan kuartil dari kolom numerik.

4. Pemeriksaan Missing Value

```
print(df.isnull().sum())
print(df[df.isnull().any(axis=1)])
```

Mengecek jumlah nilai kosong (missing values) di setiap kolom (df.isnull().sum()) dan menampilkan baris yang memiliki nilai kosong jika ada (df[df.isnull().any(axis=1)]).

5. Visualisasi Distribusi GPA

```
plt.figure(figsize=(8,5))
sns.histplot(df['GPA'], kde=True, color='skyblue')
plt.title('Distribusi GPA')
plt.xlabel('GPA')
plt.ylabel('Jumlah siswa')
plt.show()
```

Membuat histogram distribusi nilai GPA siswa dengan kurva KDE (Kernel Density Estimation) untuk menunjukkan bentuk sebarannya secara halus, membantu memahami pola umum nilai akademik siswa.

6. Boxplot GPA Berdasarkan Gender

```
plt.figure(figsize=(8,5))
sns.boxplot(x='Gender',y='GPA',data=df)
plt.title('Distribusi GPA berdasarkan Gender')
plt.xlabel('Gender(0 = laki-laki, 1 = perpmpuan)')
plt.ylabel('GPA')
plt.show()
```

Menampilkan boxplot untuk membandingkan distribusi GPA antara siswa laki-laki dan perempuan, serta mendeteksi perbedaan median, sebaran, dan outlier antar gender.

7. Heatmap Korelasi Antar Variabel

```
plt.figure(figsize=(12,8))
sns.heatmap(df.corr(),annot = True, cmap='coolwarm', fmt='.2f')
plt.title('Korelasi antar variabel')
plt.show()
```

Menghasilkan heatmap korelasi untuk menampilkan hubungan linier antar variabel numerik, membantu mengidentifikasi variabel yang paling berpengaruh terhadap GPA.

8. Scatter Plot Study Time vs GPA

```
plt.figure(figsize=(8,5))
sns.scatterplot(x='StudyTimeWeekly',y='GPA',data=df)
plt.title('Scatter Plot Study Time vs GPA')
plt.xlabel('Study Time Weekly')
plt.ylabel('GPA')
plt.show()
```

Menampilkan scatter plot antara waktu belajar mingguan dan GPA untuk melihat pola hubungan linier, yang menjadi dasar pembangunan model regresi.

9. Countplot Grade Class

```
plt.figure(figsize=(8,5))
sns.countplot(x='GradeClass', data = df, palette='viridis')
plt.title('Jumlah siswa berdasarkan Grade Class')
plt.xlabel('Grade Class')
plt.ylabel('Jumlah siswa')
plt.show()
```

Menampilkan jumlah siswa dalam tiap kategori Grade Class menggunakan countplot, membantu memahami distribusi performa akademik siswa berdasarkan kelas nilai.

10. Menampilkan Kolom yang Relevan untuk Pemodelan

```
df[['StudyTimeWeekly','Absences','Tutoring','GradeClass','GPA']].head()
```

Menampilkan beberapa baris awal dari kolom-kolom penting yang akan digunakan dalam pemodelan regresi, yaitu: StudyTimeWeekly, Absences, Tutoring, GradeClass, dan GPA.

11. Mempersiapkan Data untuk Pelatihan Model

```
X = df[['StudyTimeWeekly','Absences','Tutoring','GradeClass']]
y = df['GPA']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Memisahkan fitur prediktor (X) dan target (y), lalu membagi data menjadi data latih dan data uji menggunakan train_test_split dengan proporsi 80% untuk pelatihan dan 20% untuk pengujian.

12. Melatih Model Regresi Linier

```
model = LinearRegression()
model.fit(X_train, y_train)
```

Membuat objek model regresi linier dan melatihnya menggunakan data latih (X_train, y_train) untuk mempelajari hubungan antara variabel input dan nilai GPA.

13. Melakukan Prediksi dengan Model

```
y_pred = model.predict(X_test)
```

Memprediksi nilai GPA pada data uji (X_test) menggunakan model regresi linier yang telah dilatih sebelumnya.

14. Menampilkan Intercept dan Koefisien Model

```
print('Intercept : ', model.intercept_)
print('Koefisien : ', model.coef_)
```

Mencetak intercept (β_0) dan koefisien (β_1, β_2, dst.) dari model regresi linier, yang menunjukkan kontribusi masing-masing variabel input terhadap prediksi GPA.

15. Evaluasi Performa Model

```
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)
print ('RMSE : ', rmse)
print ('R2 : ', r2)
```

RMSE (Root Mean Squared Error): Mengukur seberapa besar kesalahan prediksi, makin kecil makin baik.
R^2 (Koefisien Determinasi): Mengukur seberapa baik model menjelaskan variansi data, nilai mendekati 1 menandakan model yang baik.

16. Visualisasi Hasil Prediksi GPA

```
plt.figure(figsize=(8,6))
plt.scatter(y_test, y_pred, alpha=0.7)
plt.xlabel('Nilai Aktual GPA')
plt.ylabel('Prediksi GPA')
plt.title('Prediksi GPA berdasarkan Study Time dan Absences')
plt.plot([y.min(),y.max()],[y.min(),y.max()],'r--')
plt.grid()
plt.show()
```

Membuat scatter plot antara nilai aktual dan nilai prediksi GPA, lalu menambahkan garis referensi (garis merah putus-putus) untuk menunjukkan prediksi sempurna. Plot ini membantu melihat seberapa dekat hasil prediksi terhadap nilai sebenarnya.

17. Fungsi Prediksi GPA Manual

```
def prediksi_gpa (studytime, absences, tutoring, gradeclass):
  intercept = 3.0512761494850533
  coef_studytime = 0.0307564
  coef_absence = -0.09918908
  coef_tutoring = 0.23301886
  coef_gradeclass = -0.11886111
  gpa = intercept + (coef_studytime * studytime) + (coef_absence * absences) + (coef_tutoring * tutoring) + (coef_gradeclass * gradeclass)
  return round (gpa,2)

studytime = float(input('Masukkan Study Time Weekly : '))
absences = float(input('Masukkan Jumlah Absences : '))
tutoring = float(input('Masukkan Jumlah Tutoring : '))
gradeclass = float(input('Masukkan Grade Class : '))
gpa_prediksi = prediksi_gpa (studytime, absences, tutoring, gradeclass)
print ('Prediksi GPA berdasarkan studytime, absences, tutoring dan gradeclass adalah : ', gpa_prediksi)
```

Membuat fungsi prediksi GPA berdasarkan input pengguna, menggunakan nilai intercept dan koefisien dari model regresi yang telah dilatih. Pengguna memasukkan nilai untuk studytime, absences, tutoring, dan gradeclass, lalu fungsi menghitung dan menampilkan prediksi GPA-nya secara langsung.

18. Hasil Prediksi

```
Masukkan Study Time Weekly : 19.833723
Masukkan Jumlah Absences : 7
Masukkan Jumlah Tutoring : 1
Masukkan Grade Class : 2.0
Prediksi GPA berdasarkan studytime, absences, tutoring dan gradeclass adalah :  2.96
```

	StudentID	Age	Gender	Ethnicity	ParentalEducation	StudyTimeWeekly	Absences	Tutoring	ParentalSupport	Extracurricular	Sports	Music	Volunteering	GPA	GradeClass
0	1001	17	1	0	2	19.833723	7	1	2	0	0	1	0	2.929196	2.0

Model regresi yang dibangun mampu memprediksi nilai GPA dengan cukup akurat, terbukti dari perbandingan antara nilai prediksi dan nilai GPA aktual yang menunjukkan selisih relatif kecil. Hal ini mengindikasikan bahwa model dapat menangkap pola akademik secara efektif dan menghasilkan estimasi GPA yang mendekati nilai sebenarnya.

3.2 Library

Library yang digunakan :

Proyek ini dibangun menggunakan bahasa pemrograman Python, dengan pustaka (library) utama sebagai berikut:

a. Pandas: digunakan untuk membaca dan memproses dataset.
b. Numpy : digunakan untuk melakukan komputasi numerik yang efisien di Python, terutama dengan array multidimensi.
c. Matplotlib.pyplot : digunakan untuk membuat berbagai jenis visualisasi data seperti grafik garis, batang, lingkaran, dan scatter.
d. Scikit-learn digunakan untuk membangun dan menerapkan berbagai model machine learning di Python, baik untuk klasifikasi, regresi, maupun klastering. Dengan modul :
    1. Model_selection : digunakan untuk membagi data, melakukan validasi silang (*cross-validation*), dan pencarian model terbaik.
    2. Linear_model : digunakan untuk membangun model regresi linier dan klasifikasi linier seperti Linear Regression.
    3. Metrics : digunakan untuk mengevaluasi kinerja model machine learning dengan berbagai metrik, seperti akurasi, precision, recall, F1-score, dan R^2.

4. Hasil dan Evaluasi

4.1 Hasil Eksperimen

Pada eksperimen ini, dilakukan analisis data (EDA) untuk memahami faktor-faktor yang memengaruhi GPA siswa. Hasil visualisasi menunjukkan bahwa StudyTimeWeekly memiliki hubungan positif dengan GPA, berarti semakin sering siswa belajar, GPA cenderung naik. Sebaliknya, Absences berkorelasi negatif dengan GPA, berarti semakin sering siswa absen, GPA cenderung turun. Selain itu, faktor Tutoring dan GradeClass juga dianalisis, berarti siswa yang mengikuti bimbingan belajar (Tutoring) atau berada di kelas tertentu (GradeClass) dapat memiliki pengaruh tersendiri terhadap pencapaian GPA mereka.

Model Multiple Linear Regression dibangun menggunakan empat variabel prediktor, yaitu StudyTimeWeekly, Absences, Tutoring, dan GradeClass. Dalam proses ini, variabel kategorikal seperti Tutoring dan GradeClass perlu dikodekan ke dalam bentuk numerik agar dapat digunakan dalam model regresi. Setelah model dilatih menggunakan data training, dilakukan evaluasi terhadap data testing menggunakan metrik Root Mean Squared Error (RMSE) dan R-squared (R^2).

Hasil evaluasi menunjukkan bahwa model dapat menjelaskan sebagian variasi nilai GPA berdasarkan keempat faktor tersebut. Ini berarti model mampu menangkap hubungan antara kombinasi faktor-faktor belajar, kehadiran, bimbingan belajar, dan kelas siswa terhadap performa GPA, walaupun masih terdapat kesalahan prediksi yang wajar dalam penerapannya. Scatter plot antara nilai aktual GPA dan hasil prediksi menunjukkan penyebaran yang cukup mendekati garis ideal, berarti model memberikan hasil yang cukup akurat dalam konteks data yang tersedia.

Model ini kemudian divisualisasikan dan diintegrasikan ke dalam dashboard interaktif menggunakan Looker Studio, memungkinkan pengguna untuk memahami hubungan antar variabel. Model ini berfungsi sebagai alat bantu analisis untuk memprediksi GPA berdasarkan data terukur, meskipun hasil prediksi tetap perlu dipertimbangkan bersama faktor eksternal lain yang tidak dianalisis dalam model ini.

4.2 Evaluasi

Evaluasi dilakukan dengan menggunakan metrik Root Mean Squared Error (RMSE) dan R-squared (R^2) untuk mengukur kinerja model setelah dilatih dengan empat variabel prediktor, yaitu StudyTimeWeekly, Absences, Tutoring, dan GradeClass. Nilai RMSE menunjukkan bahwa model memiliki tingkat kesalahan prediksi yang masih dapat diterima, berarti model mampu memberikan hasil yang cukup mendekati kenyataan walaupun tidak sempurna. Sementara itu, nilai R^2 yang berkisar sekitar 0.902 menunjukkan bahwa model mampu menjelaskan sekitar 90% variasi GPA berdasarkan variabel yang digunakan, berarti ada kontribusi signifikan dari faktor-faktor tersebut terhadap GPA. Hasil visualisasi scatter plot antara GPA aktual dan prediksi memperlihatkan bahwa model menghasilkan prediksi yang cukup akurat, berarti model ini efektif sebagai alat bantu analisis, meskipun tidak dapat dijadikan acuan tunggal karena masih ada faktor lain yang memengaruhi hasil GPA di luar model ini.

5. Kesimpulan

5.1 Kesimpulan

Model regresi yang dibangun menunjukkan bahwa variabel seperti waktu belajar mingguan, tingkat absensi, keterlibatan dalam tutoring, dan grade class memiliki kontribusi signifikan terhadap pencapaian akademik siswa yang diukur melalui GPA. Dari analisis korelasi, absensi dan durasi belajar merupakan faktor paling berpengaruh. Secara spesifik, GPA cenderung meningkat dengan bertambahnya jam belajar dan keterlibatan dalam bimbingan, namun menurun akibat tingginya absensi dan tingkat kelas yang lebih tinggi. Dengan nilai R^2 sebesar 0.902 dan RMSE sebesar 0.285, model ini dinilai akurat dan mampu menjelaskan lebih dari 90% variasi GPA. Model dikembangkan menggunakan pembagian data latih dan uji (80:20), dan hasilnya mendukung pemahaman yang lebih mendalam tentang faktor-faktor akademik dalam Students Performance Dataset.

5.2 Kontribusi

1. Vianda Retnaningtiyas Purbandari Karetji (23.11.5445) : Melakukan eksperimen di google colab, membuat poster dan laporan.
2. Rifky Danu Asmoro (23.11.5489) : Membuat Laporan, dashboard, dan melakukan eksperimen di google colab.
3. Stefanus Arya Bayu Samudra Bataona (23.11.5477) : Membuat Laporan, dashboard dan melakukan eksperimen di google colab.

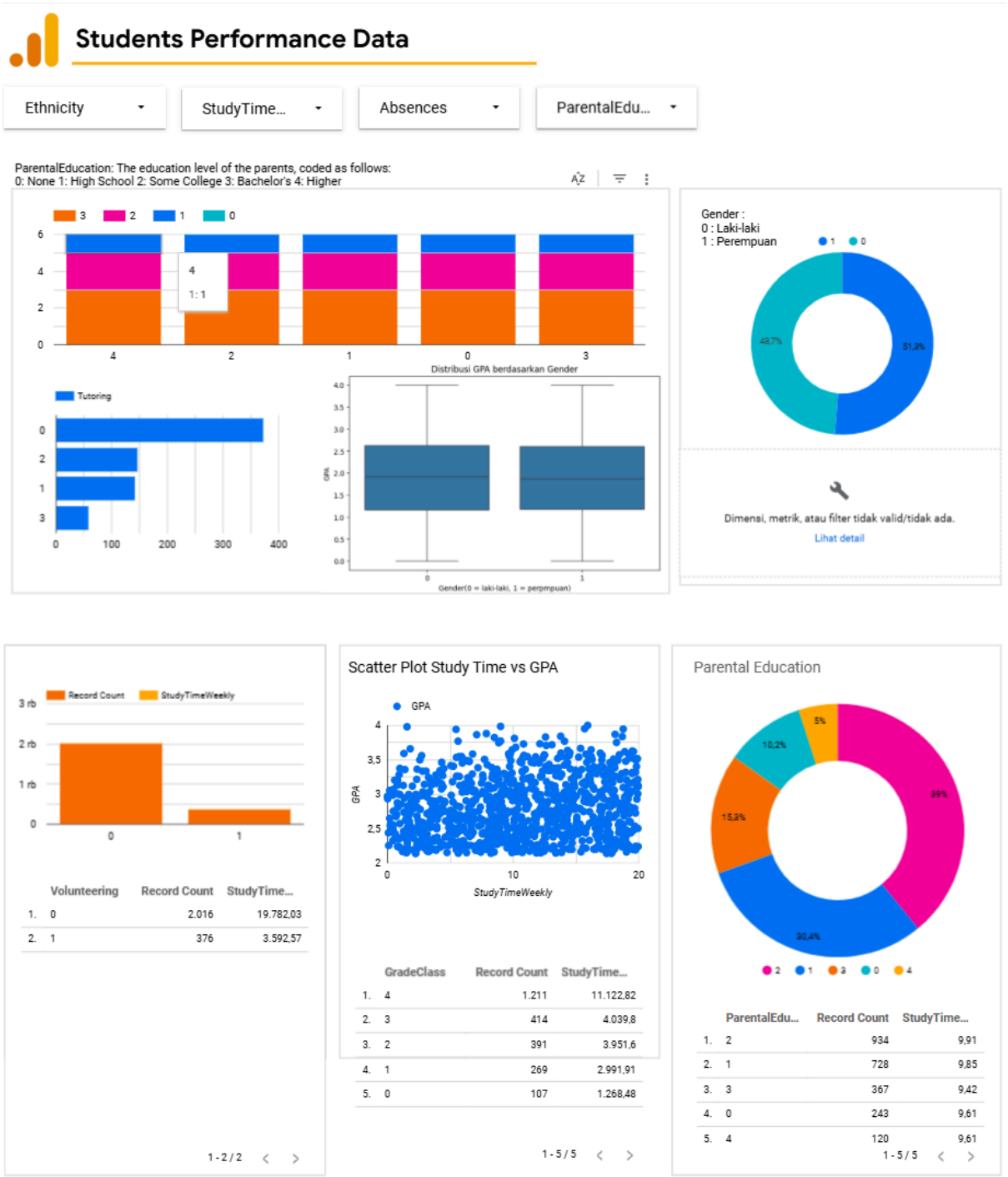
6. Lampiran

a. Link Google Colab
Prediksi GPA.ipynb
https://colab.research.google.com/drive/1T2v0a0aOTFfnUSHxRyvLFGA2hosTnn-D?usp=sharing

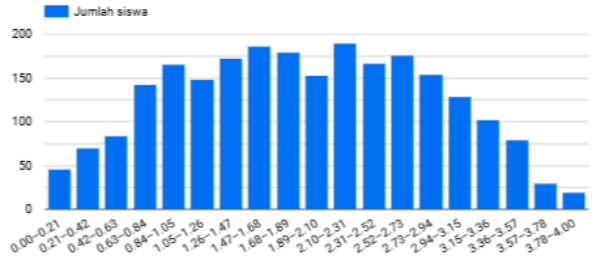
b. Link Google Drive
UAS BIG DATA & PREDICTIVE ANALYTICS
https://drive.google.com/drive/folders/1Y09WC013MCXZydNvN6_M1qcN3KFD8Dzd?usp=drive_link

c. Dataset
https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset/data

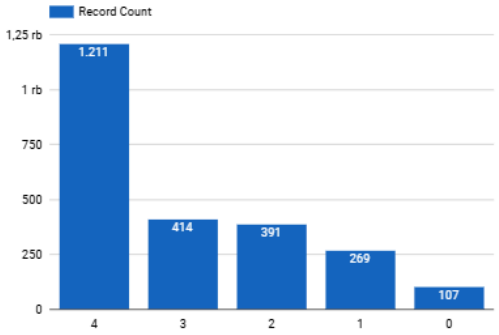
d. Dashboard
https://lookerstudio.google.com/reporting/096fe8ff-102c-4c5a-8449-96b16251d9b6

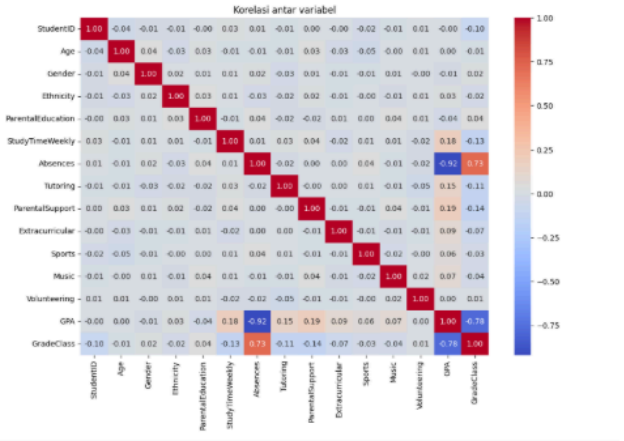


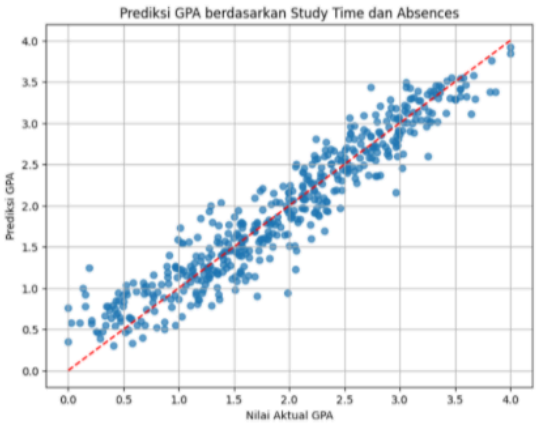
Distribusi GPA



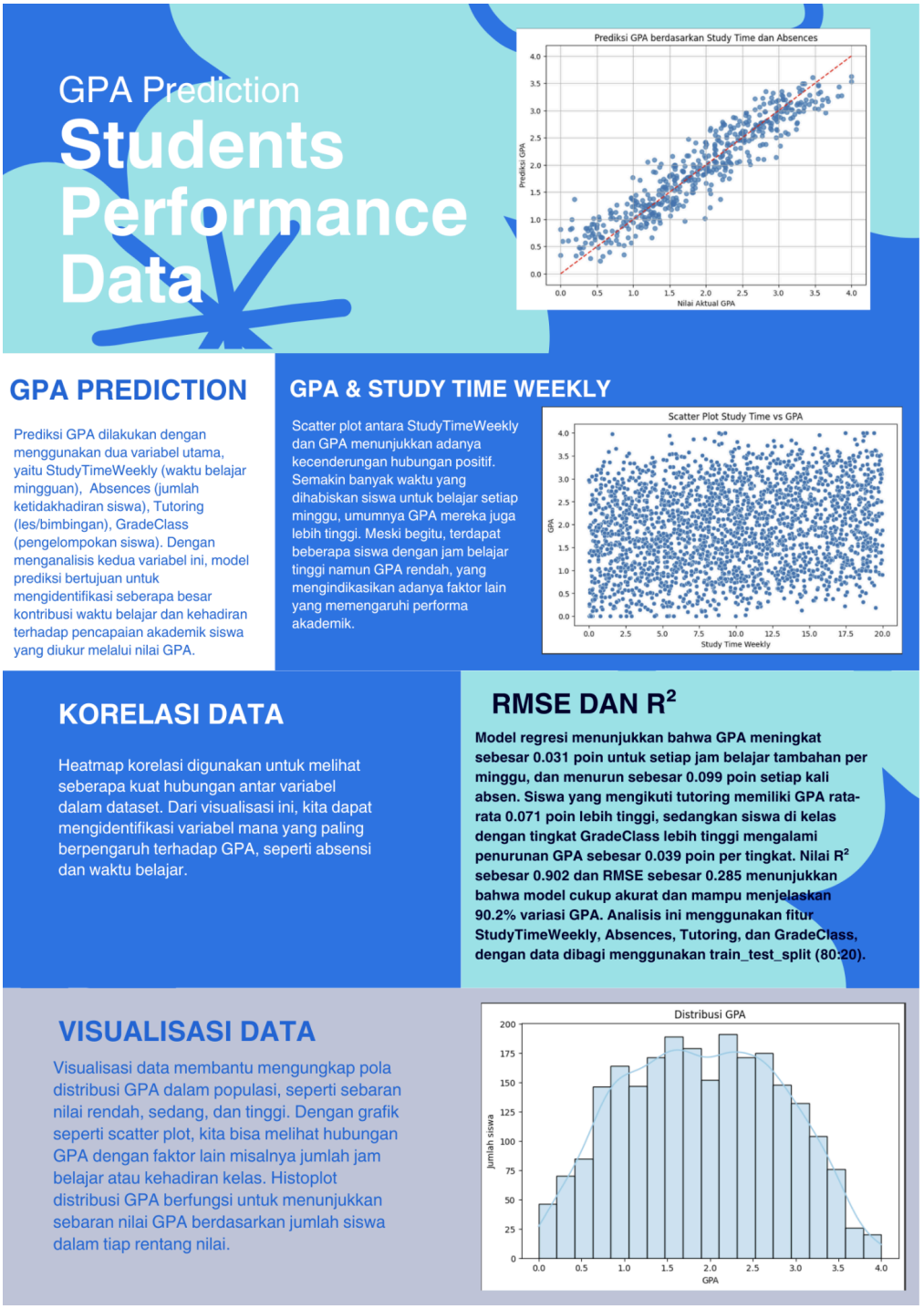
Jumlah siswa berdasarkan Grade Class







e. Poster
https://www.canva.com/design/DAGtPSUKm-U/Sk4YLKVgs23Six1M0jH_Kw/edit?utm_content=DAGtPSUKm-U&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton



GPA Prediction Dataset Explanation

STUDENTID

StudentID adalah kode unik yang diberikan kepada masing-masing siswa untuk membedakan satu sama lain dalam sistem data. Dalam rentang 1001 hingga 3392, setiap angka tersebut mewakili satu individu secara spesifik

ETHNICITY

Ethnicity merepresentasikan latar belakang etnis masing-masing siswa, yang dikodekan dalam bentuk angka untuk memudahkan analisis data. Angka 0 hingga 3 menunjukkan kelompok etnis

0 : digunakan untuk merepresentasikan siswa dengan latar belakang etnis Caucasian

1 : digunakan untuk merepresentasikan siswa dengan latar belakang etnis African-American

2 : digunakan untuk merepresentasikan siswa dengan latar belakang etnis Asian

3 : digunakan untuk merepresentasikan siswa dengan latar belakang selain itu

PARENTAL EDUCATION

ParentalEducation menunjukkan tingkat pendidikan yang telah ditempuh oleh orang tua siswa. Kode 0 hingga 4 digunakan merepresentasikan :
0 : tingkat pendidikan None (tidak disebutkan)
1 : tingkat pendidikan High School (SMA)
2 : tingkat pedidikan Some College (telah mengikuti pendidikan tinggi setelah SMA, tetapi belum menyelesaikan gelar sarjana)
3 : tingkat pedidikan Bachelor's (S1)
4: tingkat pedidikan Higher (Tingkatan diatas S1)

PARENTAL SUPPORT

ParentalSupport menggambarkan tingkat dukungan emosional, akademik, atau finansial yang diberikan orang tua kepada anaknya. Kode 0 hingga 4 digunakan untuk merepresentasikan :
0 : None (tidak ada parental support)
1 : Low (tingkat parental support rendah)
2 : Moderate (tingkat parental support rendah)
3 : High (tingkat parental support tinggi)
4 : Very High (tingkat parental support sangat tinggi)

EXTRACURRICULAR ACTIVITIES

- Extracurricular : Partisipasi dalam kegiatan ekstrakurikuler, di mana 0 berarti Tidak dan 1 berarti Ya
- Sports : Partisipasi dalam olahraga, di mana 0 berarti Tidak dan 1 berarti Ya
- Music : Partisipasi dalam kegiatan musik, di mana 0 berarti Tidak dan 1 berarti Ya
- Volunteering : Partisipasi dalam kegiatan sukarela, di mana 0 berarti Tidak dan 1 berarti Ya

STUDY HABITS

- StudyTimeWeekly : Waktu belajar siswa setiap minggu, dihitung dalam jam.
- Absences : Jumlah ketidakhadiran siswa selama tahun ajaran.
- Tutoring : Status les atau bimbingan belajar, di mana angka 0 berarti Tidak mengikuti les, dan 1 berarti Mengikuti les.