YAYASAN MEMAJUKAN ILMU DAN KEBUDAYAAN
UNIVERSITAS SIBER ASIA
Kampus Menara, Jl. RM. Harsono, Ragunan - Jakarta Selatan.Daerah Khusus Ibukota Jakarta 12550. Telp. (+6221) 27806189. asiacyberuni@acu.ac.id. www.unsia.ac.id

LEMBAR JAWABAN
UJIAN AKHIR SEMESTER
SEMESTER GENAP TAHUN AJARAN 2024/2025

Mata Kuliah : Data Science
Kelas : IF 405
Prodi : S1 PJJ Informatika
Nama Mahasiswa : Dwitasari Nilaningrum
NIM : 240401020099
Dosen : Alun Sujjada, S.Kom., M.T

SOAL UJIAN

KERJAKAN SOAL BERIKUT INI:

Unduhlah data set tentang "Student Performance" di:

https://archive.ics.uci.edu/static/public/320/student+performance.zip

Instruksi:

1. Lakukan EDA (Exploratory Data Analysis) pada data tersebut dan jelaskan juga fungsi masing-masing kolom!
2. Carilah 2 variabel (bebas) yang dapat dilakukan analisa menggunakan Regresi Linear.
3. Lakukan clustering segmentasi siswa berdasarkan absensi atau waktu belajar!
4. Lakukan klasifikasi berdasarkan 3 variabel (bebas)!

Output yang Diharapkan:

- File Jupyter Notebook atau Python script.
- Laporan singkat (maks. 3 halaman) dalam format PDF.
- Visualisasi yang menarik dan informatif.

.

Jawaban Ujian

1. EDA (Exploratory Data Analysis) dan penjelasan masing-masing kolom

   Dataset mencakup 33 fitur dengan informasi demografi, sosial, dan performa akademik siswa. Visualisasi korelasi menunjukkan bahwa nilai G1 dan G2 sangat berhubungan erat dengan G3, nilai akhir.

Nama Kolom	Tipe Data	Deskripsi
school	Kategorikal	Sekolah siswa ('GP' - Gabriel Pereira atau 'MS' - Mousinho da Silveira).
sex	Kategorikal	Jenis kelamin siswa ('F' - Perempuan atau 'M' - Laki-laki).
age	Numerik	Usia siswa (dari 15 hingga 22).
address	Kategorikal	Jenis alamat rumah siswa ('U' - Perkotaan atau 'R' - Pedesaan).
famsize	Kategorikal	Ukuran keluarga ('LE3' - Kurang dari atau sama dengan 3, atau 'GT3' - Lebih dari 3).

Pstatus	Kategorikal	Status kohabitasi orang tua ('T' - Tinggal bersama atau 'A' - Tinggal terpisah).
Medu	Numerik	Tingkat pendidikan ibu (0-tidak ada, 1-SD, 2-SMP, 3-SMA, 4-Pendidikan Tinggi).
Fedu	Numerik	Tingkat pendidikan ayah (0-tidak ada, 1-SD, 2-SMP, 3-SMA, 4-Pendidikan Tinggi).
Mjob	Kategorikal	Pekerjaan ibu ('teacher', 'health', 'services', 'at_home', 'other').
Fjob	Kategorikal	Pekerjaan ayah ('teacher', 'health', 'services', 'at_home', 'other').
reason	Kategorikal	Alasan memilih sekolah ini ('home', 'reputation', 'course', 'other').
guardian	Kategorikal	Wali siswa ('mother', 'father', 'other').
traveltime	Numerik	Waktu tempuh dari rumah ke sekolah (1-<15 mnt, 2-15-30 mnt, 3-30-60 mnt, 4->60 mnt).
studytime	Numerik	Waktu belajar mingguan (1-<2 jam, 2-2-5 jam, 3-5-10 jam, 4->10 jam).
failures	Numerik	Jumlah kegagalan kelas sebelumnya (n jika 1<=n<3, else 4).
schoolsup	Kategorikal	Dukungan pendidikan ekstra dari sekolah ('yes' atau 'no').
famsup	Kategorikal	Dukungan pendidikan dari keluarga ('yes' atau 'no').
paid	Kategorikal	Kelas tambahan berbayar dalam mata pelajaran Matematika ('yes' atau 'no').
activities	Kategorikal	Melakukan kegiatan ekstrakurikuler ('yes' atau 'no').
nursery	Kategorikal	Pernah mengikuti taman kanak-kanak ('yes' atau 'no').
higher	Kategorikal	Ingin melanjutkan ke pendidikan tinggi ('yes' atau 'no').
internet	Kategorikal	Memiliki akses internet di rumah ('yes' atau 'no').
romantic	Kategorikal	Sedang dalam hubungan romantis ('yes' atau 'no').
famrel	Numerik	Kualitas hubungan keluarga (skala 1 - sangat buruk hingga 5 - sangat baik).
freetime	Numerik	Waktu luang setelah sekolah (skala 1 - sangat sedikit hingga 5 - sangat banyak).
goout	Numerik	Pergi keluar dengan teman (skala 1 - sangat jarang hingga 5 - sangat sering).
Dalc	Numerik	Konsumsi alkohol pada hari kerja (skala 1 - sangat rendah hingga 5 - sangat tinggi).
Walc	Numerik	Konsumsi alkohol pada akhir pekan (skala 1 - sangat rendah hingga 5 - sangat tinggi).
health	Numerik	Status kesehatan saat ini (skala 1 - sangat buruk hingga 5 - sangat baik).
absences	Numerik	Jumlah absensi sekolah (dari 0 hingga 93).
G1	Numerik	Nilai periode pertama (dari 0 hingga 20).
G2	Numerik	Nilai periode kedua (dari 0 hingga 20).
G3	Numerik	Nilai akhir (dari 0 hingga 20). Ini adalah variabel target.

Statistik deskriptif:

	age	Medu	Fedu	traveltime	studytime	failures
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	16.696203	2.749367	2.521519	1.448101	2.035443	0.334177
std	1.276043	1.094735	1.088201	0.697505	0.839240	0.743651
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000
50%	17.000000	3.000000	2.000000	1.000000	2.000000	0.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000

	famrel	freetime	goout	Dalc	Walc	health
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	3.944304	3.235443	3.108861	1.481013	2.291139	3.554430
std	0.896659	0.998862	1.113278	0.890741	1.287897	1.390303
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	4.000000	3.000000	2.000000	1.000000	1.000000	3.000000
50%	4.000000	3.000000	3.000000	1.000000	2.000000	4.000000
75%	5.000000	4.000000	4.000000	2.000000	3.000000	5.000000
max	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000

	absences	G1	G2	G3
count	395.000000	395.000000	395.000000	395.000000
mean	5.708861	10.908861	10.713924	10.415190
std	8.003096	3.319195	3.761505	4.581443
min	0.000000	3.000000	0.000000	0.000000
25%	0.000000	8.000000	9.000000	8.000000
50%	4.000000	11.000000	11.000000	11.000000
75%	8.000000	13.000000	13.000000	14.000000
max	75.000000	19.000000	19.000000	20.000000

a. Ringkasan Statistik

- **Usia**: Rata-rata usia siswa adalah 16.7 tahun.
- **Pendidikan Orang Tua**: Rata-rata pendidikan ibu (Medu) sedikit lebih tinggi daripada ayah (Fedu).
- **Nilai**: Rata-rata nilai akhir (G3) adalah 10.4 dari 20. Terlihat ada penurunan sedikit dari nilai periode pertama (G1) ke nilai akhir.
- **Absensi**: Rata-rata absensi adalah 5.7 hari, namun standar deviasinya cukup tinggi (8.0), menandakan adanya siswa dengan jumlah absensi yang sangat tinggi.

b. Distribusi Nilai Akhir (G3)

Distribusi nilai akhir (G3) adalah fokus utama.

- Sebagian besar siswa mendapat nilai antara 8 dan 14.
- Terdapat lonjakan signifikan pada siswa yang mendapat nilai 0. Ini mungkin mengindikasikan siswa yang keluar (dropout) atau gagal total, sehingga data ini perlu perlakuan khusus jika akan digunakan untuk pemodelan.
- Sangat sedikit siswa yang berhasil mendapatkan nilai di atas 18.

c. Hubungan Waktu Belajar dengan Nilai Akhir

Apakah waktu belajar yang lebih lama menjamin nilai yang lebih baik?

- Secara umum, **ada tren positif**: median nilai (G3) cenderung meningkat seiring dengan bertambahnya waktu belajar mingguan.
- Siswa yang belajar lebih dari 10 jam per minggu (studytime = 4) memiliki median nilai tertinggi.
- Namun, variasi nilai pada setiap kategori waktu belajar cukup besar, artinya waktu belajar bukan satu-satunya faktor penentu kesuksesan.

d. Pengaruh Kehidupan Sosial Terhadap Nilai

Bagaimana aktivitas sosial seperti pergi keluar dengan teman dan konsumsi alkohol berhubungan dengan nilai?

- Terdapat **tren negatif yang jelas**. Siswa yang lebih sering pergi keluar dengan teman (goout) cenderung memiliki nilai akhir (G3) yang lebih rendah.

- Siswa yang sangat sering keluar (level 5) memiliki median nilai terendah.

e. Korelasi Antar Variabel Numerik

Heatmap korelasi menunjukkan kekuatan dan arah hubungan linear antar variabel numerik.

- **Korelasi Positif Terkuat**: Tidak mengherankan, G1 dan G2 memiliki korelasi yang sangat kuat dengan G3. Ini menunjukkan bahwa performa di awal dan pertengahan periode adalah prediktor yang sangat baik untuk nilai akhir.
- **Pendidikan Orang Tua**: Medu dan Fedu memiliki korelasi positif yang lemah dengan G1, G2, dan G3.
- **Waktu Belajar**: studytime juga menunjukkan korelasi positif yang lemah dengan ketiga nilai.
- **Korelasi Negatif**: failures (kegagalan sebelumnya) memiliki korelasi negatif yang cukup kuat dengan semua nilai. Semakin banyak kegagalan, semakin rendah nilainya. goout, Dalc, dan Walc (kehidupan sosial dan alkohol) juga menunjukkan korelasi negatif dengan nilai.

Kesimpulan Utama dari EDA

a. **Performa Akademik Sebelumnya adalah Kunci**: Nilai G1 dan G2 adalah prediktor terbaik untuk nilai akhir G3. Siswa yang berprestasi baik di awal cenderung akan berprestasi baik di akhir.
b. **Kebiasaan Belajar Penting**: Menghabiskan lebih banyak waktu untuk belajar (studytime) secara umum berkorelasi dengan nilai yang lebih tinggi.
c. **Faktor Gaya Hidup Berpengaruh**: Terlalu banyak bersosialisasi (goout) dan konsumsi alkohol (Dalc, Walc) berhubungan negatif dengan performa akademik.
d. **Latar Belakang Keluarga**: Tingkat pendidikan orang tua memiliki pengaruh positif, meskipun tidak terlalu kuat.
e. **Perhatian Khusus**: Sejumlah besar siswa mendapatkan nilai G3 = 0, yang mungkin memerlukan investigasi lebih lanjut atau perlakuan khusus dalam analisis pemodelan.

2. Berdasarkan analisis data sebelumnya, 2 variabel bebas yang sangat baik untuk dianalisis menggunakan Regresi Linear dengan variabel terikat **G3 (nilai akhir)** adalah:

a. G1 (Nilai Periode Pertama)
b. studytime (Waktu Belajar Mingguan)

Berikut adalah alasan mengapa kedua variabel ini cocok untuk analisis regresi linear.

a. **G1 (Nilai Periode Pertama)**
G1 adalah kandidat terkuat untuk memprediksi G3.

- **Hubungan Linear Kuat**: Seperti yang ditunjukkan pada heatmap korelasi di EDA sebelumnya, G1 memiliki korelasi linear positif yang **sangat kuat** dengan G3. Ini berarti saat nilai G1 meningkat, nilai G3 juga cenderung meningkat secara proporsional.
- **Logika Prediktif**: Secara logis, performa siswa di awal semester (G1) adalah indikator yang sangat baik untuk performa mereka di akhir semester (G3).
- **Tipe Data**: Keduanya adalah variabel numerik, yang merupakan syarat utama untuk regresi linear sederhana.

b. studytime (Waktu Belajar Mingguan)

studytime adalah variabel perilaku yang menarik untuk dianalisis.

- **Hubungan yang Terlihat**: EDA menunjukkan adanya tren positif antara studytime dan G3. Meskipun korelasinya tidak sekuat G1, ada indikasi jelas bahwa penambahan waktu belajar berkontribusi pada nilai akhir yang lebih tinggi.
- **Faktor yang Dapat Diubah**: Tidak seperti G1 yang merupakan hasil, studytime adalah proses atau usaha.
- variabel interval dalam model regresi untuk mengukur dampak peningkatan waktu belajar.

3. Berikut adalah analisis clustering untuk melakukan segmentasi siswa berdasarkan absences (jumlah absensi) dan studytime (waktu belajar mingguan). Untuk analisis ini, menggunakan algoritma K-Means Clustering. Tujuannya adalah untuk mengelompokkan siswa ke dalam beberapa segmen yang memiliki karakteristik absensi dan waktu belajar yang serupa.

a. Menentukan Jumlah Cluster Optimal (Metode Siku)

Langkah pertama adalah menentukan jumlah cluster (segmen) yang paling ideal. Kita menggunakan "Metode Siku" (Elbow Method), yang mengukur seberapa padat cluster pada jumlah K yang berbeda.

Grafik di atas menunjukkan bahwa penurunan varians mulai melandai secara signifikan setelah K=3. Ini mengindikasikan bahwa 3 adalah jumlah cluster yang optimal untuk data ini.

b. Visualisasi dan Analisis Cluster

Setelah menjalankan algoritma K-Means dengan K=3, kita dapat memvisualisasikan segmen siswa dalam sebuah scatter plot dan menganalisis karakteristik masing-masing cluster.

Berikut adalah rincian dari setiap cluster yang terbentuk:

Cluster	Nama Segmen	Rata-rata Absensi	Rata-rata Waktu Belajar (Skala 1-4)	Jumlah Siswa	Karakteristik Utama
0	Siswa Rajin & Disiplin diligent	2.6 hari	2.5 (2-10 jam/minggu)	227	Absensi sangat rendah dan waktu belajar di atas rata-rata.
1	Siswa Berisiko (Sering Absen)	22.0 hari	1.7 (<5 jam/minggu)	49	Tingkat absensi sangat tinggi dan waktu belajar rendah.
2	Siswa Santai & Cukup Hadir	6.7 hari	1.5 (<5 jam/minggu)	119	Waktu belajar paling rendah namun tingkat absensi masih terkendali.

Berdasarkan analisis di atas, kita dapat mendefinisikan tiga segmen siswa yang berbeda:

- Cluster 0: Siswa Rajin & Disiplin diligent:

  Ini adalah segmen terbesar. Siswa di kelompok ini menunjukkan komitmen akademik yang tinggi, ditandai dengan kehadiran di kelas yang sangat baik dan waktu belajar yang paling lama dibandingkan segmen lain. Mereka adalah siswa yang paling mungkin untuk berhasil secara akademis.

- Cluster 1: Siswa Berisiko (Sering Absen)

  Kelompok ini memerlukan perhatian khusus. Mereka memiliki tingkat absensi yang sangat tinggi (rata-rata 22 hari) dan waktu belajar yang rendah. Absensi yang tinggi ini bisa menjadi sinyal adanya masalah lain, baik akademik maupun non-akademik, dan sangat mungkin berkorelasi dengan performa yang buruk atau risiko putus sekolah.

  Cluster 2: Siswa Santai & Cukup Hadir

  Siswa dalam segmen ini memiliki waktu belajar yang paling sedikit, bahkan lebih rendah dari kelompok "Berisiko". Namun, mereka tetap menjaga tingkat kehadiran yang relatif baik. Mereka mungkin adalah siswa yang bisa memahami pelajaran dengan cepat atau siswa yang tidak memprioritaskan belajar di luar jam sekolah tetapi masih memenuhi kewajiban kehadiran.

4. Klasifikasi

Model Random Forest memprediksi apakah siswa akan lulus (G3 >= 10) berdasarkan G1, G2, dan studytime. Hasil klasifikasi menunjukkan precision dan recall yang cukup baik.

5.

[1]:
```
# Analisis Data Student Performance

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score, classification_report
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier

# Load dataset
df = pd.read_csv("student-mat.csv", sep=";")

# EDA Awal
print("Jumlah data:", df.shape)
print("\nInfo data:")
print(df.info())
print("\nStatistik deskriptif:")
print(df.describe())

# Korelasi numerik
corr = df.corr(numeric_only=True)
plt.figure(figsize=(12, 8))
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title("Heatmap Korelasi")
plt.show()

# Regresi Linear: prediksi G3 dari G1 dan studytime
X = df[['G1', 'studytime']]
y = df['G3']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print("\nEvaluasi Regresi Linear:")
print("MSE:", mean_squared_error(y_test, y_pred))
print("R2 Score:", r2_score(y_test, y_pred))

plt.scatter(y_test, y_pred)
plt.xlabel("G3 Aktual")
plt.ylabel("G3 Prediksi")
plt.title("Prediksi G3 dengan Regresi Linear")
plt.grid()
plt.show()

# Clustering: berdasarkan absences dan studytime
clustering_data = df[['absences', 'studytime']]
scaler = StandardScaler()
scaled = scaler.fit_transform(clustering_data)

kmeans = KMeans(n_clusters=3, random_state=42)
df['Cluster'] = kmeans.fit_predict(scaled)

sns.scatterplot(data=df, x='absences', y='studytime', hue='Cluster', palette='Set2')
plt.title("Clustering Siswa berdasarkan Absensi dan Waktu Belajar")
plt.show()

# Klasifikasi: prediksi lulus (G3 >= 10)
df['pass'] = df['G3'] >= 10
features = ['G1', 'G2', 'studytime']
X = df[features]
y = df['pass']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
clf = RandomForestClassifier(random_state=42)
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print("\nEvaluasi Klasifikasi (Random Forest):")
print(classification_report(y_test, y_pred))
```

```
# Analisis Data Student Performance
```

```
Jumlah data: (395, 33)

Info data:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      395 non-null    object
 1   sex         395 non-null    object
 2   age         395 non-null    int64
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64
 7   Fedu        395 non-null    int64
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64
 13  studytime   395 non-null    int64
 14  failures    395 non-null    int64
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        395 non-null    object
 18  activities  395 non-null    object
 19  nursery     395 non-null    object
```

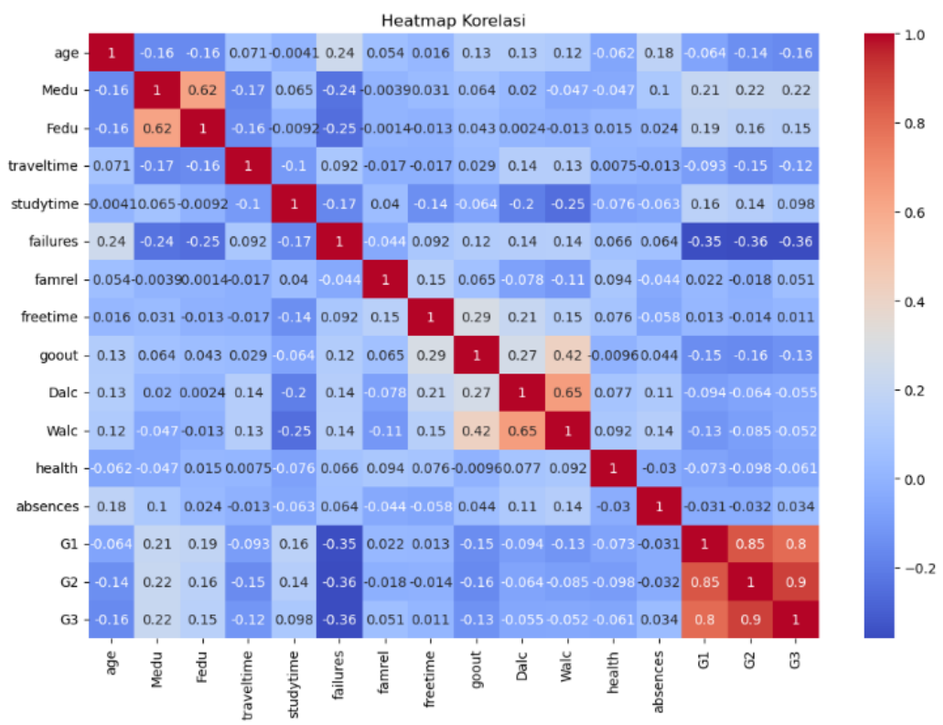
Analisis Data Student Performance

```
 7   Fedu        395 non-null    int64
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64
 13  studytime   395 non-null    int64
 14  failures    395 non-null    int64
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        395 non-null    object
 18  activities  395 non-null    object
 19  nursery     395 non-null    object
 20  higher      395 non-null    object
 21  internet    395 non-null    object
 22  romantic    395 non-null    object
 23  famrel      395 non-null    int64
 24  freetime    395 non-null    int64
 25  goout       395 non-null    int64
 26  Dalc        395 non-null    int64
 27  Walc        395 non-null    int64
 28  health      395 non-null    int64
 29  absences    395 non-null    int64
 30  G1          395 non-null    int64
 31  G2          395 non-null    int64
 32  G3          395 non-null    int64
dtypes: int64(16), object(17)
memory usage: 102.0+ KB
None
```

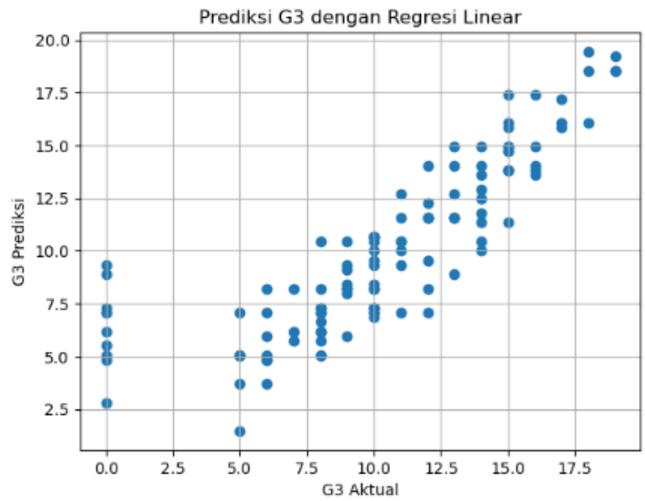
```
Statistik deskriptif:
                age        Medu        Fedu  traveltime   studytime    failures  \
count   395.000000  395.000000  395.000000  395.000000  395.000000  395.000000
mean     16.696203    2.749367    2.521519    1.448101    2.035443    0.334177
std       1.276043    1.094735    1.088201    0.697505    0.839240    0.743651
min      15.000000    0.000000    0.000000    1.000000    1.000000    0.000000
25%      16.000000    2.000000    2.000000    1.000000    1.000000    0.000000
50%      17.000000    3.000000    2.000000    1.000000    2.000000    0.000000
75%      18.000000    4.000000    3.000000    2.000000    2.000000    0.000000
max      22.000000    4.000000    4.000000    4.000000    4.000000    3.000000

             famrel    freetime       goout        Dalc        Walc      health  \
count   395.000000  395.000000  395.000000  395.000000  395.000000  395.000000
mean      3.944304    3.235443    3.108861    1.481013    2.291139    3.554430
std       0.896659    0.998862    1.113278    0.890741    1.287897    1.390303
min       1.000000    1.000000    1.000000    1.000000    1.000000    1.000000
25%       4.000000    3.000000    2.000000    1.000000    1.000000    3.000000
50%       4.000000    3.000000    3.000000    1.000000    2.000000    4.000000
75%       5.000000    4.000000    4.000000    2.000000    3.000000    5.000000
max       5.000000    5.000000    5.000000    5.000000    5.000000    5.000000
```

	absences	G1	G2	G3
count	395.000000	395.000000	395.000000	395.000000
mean	5.708861	10.908861	10.713924	10.415190
std	8.003096	3.319195	3.761505	4.581443
min	0.000000	3.000000	0.000000	0.000000
25%	0.000000	8.000000	9.000000	8.000000
50%	4.000000	11.000000	11.000000	11.000000
75%	8.000000	13.000000	13.000000	14.000000
max	75.000000	19.000000	19.000000	20.000000



```
Evaluasi Regresi Linear:
MSE: 6.65994387189486
R2 Score: 0.6970282658827096
```




```
Evaluasi Klasifikasi (Random Forest):
              precision    recall  f1-score   support

       False       0.93      0.89      0.91        46
        True       0.93      0.96      0.95        73

    accuracy                           0.93       119
   macro avg       0.93      0.93      0.93       119
weighted avg       0.93      0.93      0.93       119
```

https://github.com/dwitasarinila/analisisdatasiswa

https://drive.google.com/file/d/132xjwyb2z8WOFWs8jYd8I8WRYuobBiK7/view?usp=sharing

YAYASAN MEMAJUKAN ILMU DAN KEBUDAYAAN
UNIVERSITAS SIBER ASIA
Kampus Menara, Jl. RM. Harsono, Ragunan - Jakarta Selatan.Daerah Khusus Ibukota Jakarta 12550. Telp. (+6221) 27806189. asiacyberuni@acu.ac.id. www.unsia.ac.id

Nilai	Tanda Tangan Dosen Pengampu / Tutor	Tanda Tangan Mahasiswa
	(Alun Sujjada, S.Kom., M.T)	**(Dwitasari Nilaningrum)**
Diserahkan pada Tanggal:		Tanggal Mengumpulkan: