R Studio Dataset
Student Health & Attendance Data

Data Analyst Camp Ousean Group

Muhammad Aziz Habiburrahim

IMPORT LIBRARY

INPUT

```
1
2  # Import library
3  library(ggplot2)
4  library(caret)
5  library(dplyr)
6
```

LOAD DATASET

INPUT

```
5
6  # Load dataset
7  library(readxl)
8  data <- read_excel("D:/Course/Ousean/data.xlsx")
9
```

LOAD DATASET

OUTPUT

	student_id	date	class_time	attendance_status	stress_level	sleep_hours	anxiety_level	mood_score	risk_level
1	1	2024-12-01	9:00-15:00	Late	0.92	7.6	6	6	Low
2	1	2024-12-02	8:00-16:00	Late	1.17	6.0	6	2	Medium
3	1	2024-12-03	11:00-14:00	Late	4.56	6.3	4	8	High
4	1	2024-12-04	11:00-16:00	Late	3.07	9.0	2	10	Low
5	1	2024-12-05	9:00-13:00	Absent	3.93	7.4	9	4	High
6	1	2024-12-06	8:00-14:00	Present	4.96	6.6	5	9	High
7	1	2024-12-07	11:00-15:00	Absent	2.93	6.8	4	5	High
8	1	2024-12-08	8:00-15:00	Absent	2.17	8.4	9	9	High
9	1	2024-12-09	11:00-13:00	Absent	4.40	5.9	4	4	High
10	1	2024-12-10	9:00-16:00	Late	1.44	7.7	3	7	Low
11	1	2024-12-11	10:00-16:00	Present	3.79	6.2	5	4	High
12	1	2024-12-12	8:00-12:00	Absent	4.82	8.3	1	8	High
13	1	2024-12-13	8:00-12:00	Late	3.29	5.0	2	2	Medium
14	1	2024-12-14	11:00-12:00	Present	2.87	7.9	1	9	Low
15	1	2024-12-15	10:00-16:00	Absent	3.37	8.3	10	10	High
16	1	2024-12-16	10:00-16:00	Present	4.48	6.5	2	9	High
17	1	2024-12-17	9:00-14:00	Present	2.92	6.0	8	1	Medium
18	1	2024-12-18	8:00-15:00	Present	3.20	5.8	1	3	Medium
19	1	2024-12-19	11:00-13:00	Present	2.35	6.6	5	4	Low
20	1	2024-12-20	10:00-13:00	Late	4.09	7.2	6	4	High
21	1	2024-12-21	8:00-14:00	Absent	2.92	5.5	9	1	High

SPLIT DATA

INPUT

```
 9
10  # Split data menjadi training dan testing
11  set.seed(123)
12  trainIndex <- createDataPartition(data$stress_level, p = 0.8, list = FALSE)
13  trainData <- data[trainIndex, ]
14  testData <- data[-trainIndex, ]
15
```

SPLIT DATA OUTPUT

	student_id	date	class_time	attendance_status	stress_level	sleep_hours	anxiety_level	mood_score	risk_level
1	1	2024-12-01	9:00-15:00	Late	0.92	7.6	6	6	Low
2	1	2024-12-02	8:00-16:00	Late	1.17	6.0	6	2	Medium
3	1	2024-12-03	11:00-14:00	Late	4.56	6.3	4	8	High
4	1	2024-12-04	11:00-16:00	Late	3.07	9.0	2	10	Low
5	1	2024-12-05	9:00-13:00	Absent	3.93	7.4	9	4	High
6	1	2024-12-06	8:00-14:00	Present	4.96	6.6	5	9	High
7	1	2024-12-07	11:00-15:00	Absent	2.93	6.8	4	5	High
8	1	2024-12-08	8:00-15:00	Absent	2.17	8.4	9	9	High
9	1	2024-12-09	11:00-13:00	Absent	4.40	5.9	4	4	High
10	1	2024-12-11	10:00-16:00	Present	3.79	6.2	5	4	High
11	1	2024-12-12	8:00-12:00	Absent	4.82	8.3	1	8	High
12	1	2024-12-13	8:00-12:00	Late	3.29	5.0	2	2	Medium
13	1	2024-12-14	11:00-12:00	Present	2.87	7.9	1	9	Low
14	1	2024-12-15	10:00-16:00	Absent	3.37	8.3	10	10	High
15	1	2024-12-16	10:00-16:00	Present	4.48	6.5	2	9	High
16	1	2024-12-17	9:00-14:00	Present	2.92	6.0	8	1	Medium
17	1	2024-12-18	8:00-15:00	Present	3.20	5.8	1	3	Medium
18	1	2024-12-19	11:00-13:00	Present	2.35	6.6	5	4	Low
19	1	2024-12-21	8:00-14:00	Absent	2.92	5.5	9	1	High
20	1	2024-12-22	9:00-12:00	Late	3.29	7.8	3	9	Low
21	1	2024-12-23	9:00-14:00	Present	1.34	7.8	7	7	Low

	student_id	date	class_time	attendance_status	stress_level	sleep_hours	anxiety_level	mood_score	risk_level
1	1	2024-12-10	9:00-16:00	Late	1.44	7.7	3	7	Low
2	1	2024-12-20	10:00-13:00	Late	4.09	7.2	6	4	High
3	1	2024-12-27	11:00-12:00	Absent	1.89	7.0	10	6	High
4	1	2024-12-28	8:00-13:00	Present	0.72	5.5	1	8	Low
5	1	2024-12-29	11:00-12:00	Late	4.94	6.6	2	3	High
6	2	2024-12-11	8:00-15:00	Late	4.58	6.2	2	6	High
7	2	2024-12-15	10:00-16:00	Present	2.16	8.5	4	7	Low
8	2	2024-12-20	8:00-14:00	Absent	4.84	7.1	7	6	High
9	2	2024-12-27	11:00-13:00	Late	2.67	5.4	4	7	Low
10	2	2024-12-28	11:00-13:00	Present	1.06	5.9	7	9	Low
11	3	2024-12-01	10:00-16:00	Present	2.56	7.2	8	9	Medium
12	3	2024-12-02	9:00-13:00	Late	2.60	6.0	4	6	Low
13	3	2024-12-08	9:00-15:00	Absent	2.16	8.1	3	1	High
14	3	2024-12-11	8:00-15:00	Late	4.66	7.8	8	7	High
15	3	2024-12-22	9:00-12:00	Present	4.11	7.2	7	8	High
16	3	2024-12-25	8:00-14:00	Absent	2.69	9.0	1	6	High
17	3	2024-12-28	11:00-16:00	Absent	4.16	6.1	10	4	High
18	4	2024-12-03	9:00-12:00	Late	1.82	6.0	4	9	Low
19	4	2024-12-04	8:00-12:00	Late	0.67	8.5	10	8	Medium
20	4	2024-12-10	10:00-16:00	Present	2.75	8.0	4	6	Low
21	4	2024-12-11	9:00-16:00	Late	3.56	5.2	1	3	High

MODEL THREE INPUT

```
15
16  # 1. Membuat tiga model regresi
17  model1 <- lm(student_id ~ stress_level + sleep_hours, data = trainData)
18  model2 <- lm(student_id ~ stress_level + sleep_hours + anxiety_level, data = trainData)
19  model3 <- lm(student_id ~ stress_level + sleep_hours + anxiety_level + mood_score, data = trainData)
20
```

MODEL THREE OUTPUT

model3	list [12] (S3: lm)	List of length 12
coefficients	double [5]	255.494 -0.410 -0.700 -0.451 0.724
residuals	double [12002]	-250 -249 -252 -253 -247 -252 ...
effects	double [12002]	-27488.6 -64.6 -91.4 -144.9 227.8 -255.0 ...
rank	integer [1]	5
fitted.values	double [12002]	251 250 253 254 248 253 ...
assign	integer [5]	0 1 2 3 4
qr	list [5] (S3: qr)	List of length 5
df.residual	integer [1]	11997
xlevels	list [0]	List of length 0
call	language	lm(formula = student_id ~ stress_level + sleep_hours + anxiety_level + mood .
terms	formula	student_id ~ stress_level + sleep_hours + anxiety_level + mood_score
model	list [12002 x 5] (S3: data.frame)	A data.frame with 12002 rows and 5 columns

model1	list [12] (S3: lm)	List of length 12
coefficients	double [3]	257.219 -0.448 -0.724
residuals	double [12002]	-250 -251 -250 -248 -249 -249 ...
effects	double [12002]	-27488.6 -64.6 -91.4 -250.6 -249.6 -249.9 ...
rank	integer [1]	3
fitted.values	double [12002]	251 252 251 249 250 250 ...
assign	integer [3]	0 1 2
qr	list [5] (S3: qr)	List of length 5
df.residual	integer [1]	11999
xlevels	list [0]	List of length 0
call	language	lm(formula = student_id ~ stress_level + sleep_hours, data = trainData)
terms	formula	student_id ~ stress_level + sleep_hours
model	list [12002 x 3] (S3: data.frame)	A data.frame with 12002 rows and 3 columns

model2	list [12] (S3: lm)	List of length 12
coefficients	double [4]	259.830 -0.433 -0.737 -0.462
residuals	double [12002]	-250 -251 -250 -250 -248 -250 ...
effects	double [12002]	-27488.6 -64.6 -91.4 -144.9 -250.8 -249.7 ...
rank	integer [1]	4
fitted.values	double [12002]	251 252 251 251 249 251 ...
assign	integer [4]	0 1 2 3
qr	list [5] (S3: qr)	List of length 5
df.residual	integer [1]	11998
xlevels	list [0]	List of length 0
call	language	lm(formula = student_id ~ stress_level + sleep_hours + anxiety_level, data ...
terms	formula	student_id ~ stress_level + sleep_hours + anxiety_level
model	list [12002 x 4] (S3: data.frame)	A data.frame with 12002 rows and 4 columns

R2 & RMSE FUNCTION

INPUT

```
20
21  # 2. Fungsi untuk menghitung R^2 dan RMSE
22  calculate_metrics <- function(model, testData) {
23    predictions <- predict(model, testData)
24    actuals <- testData$student_id
25    r2 <- cor(predictions, actuals)^2
26    rmse <- sqrt(mean((predictions - actuals)^2))
27    return(data.frame(R2 = r2, RMSE = rmse))
28  }
29
```

CALCULATE METRICS

INPUT

```r

# 3. Menghitung metrics untuk setiap model
metrics1 <- calculate_metrics(model1, testData)
metrics2 <- calculate_metrics(model2, testData)
metrics3 <- calculate_metrics(model3, testData)

metrics_df <- data.frame(
  Model = c("Model 1", "Model 2", "Model 3"),
  R2 = c(metrics1$R2, metrics2$R2, metrics3$R2),
  RMSE = c(metrics1$RMSE, metrics2$RMSE, metrics3$RMSE)
)

```

CALCULATE METRICS OUTPUT

	R2	RMSE
1	0.0009687424	144.6233

METRICS 1

	R2	RMSE
1	1.084397e-05	144.6552

METRICS 2

	R2	RMSE
1	0.0003723718	144.7258

METRICS 3

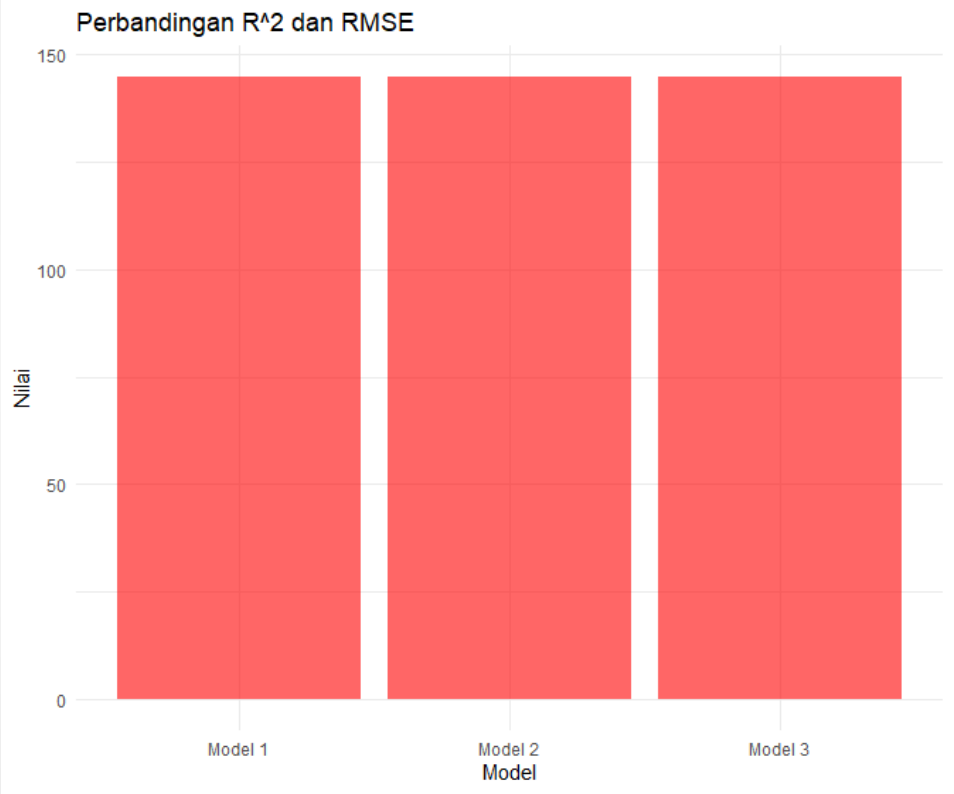
VISUALIZATION METRICS

INPUT

```
40
41  # 4. Visualisasi perbandingan metrics
42  ggplot(metrics_df, aes(x = Model)) +
43    geom_bar(aes(y = R2), stat = "identity", fill = "blue", alpha = 0.6) +
44    geom_bar(aes(y = RMSE), stat = "identity", fill = "red", alpha = 0.6) +
45    labs(title = "Perbandingan R^2 dan RMSE", y = "Nilai") +
46    theme_minimal()
47
```

METRICS 3

VISUALIZATION METRICS OUTPUT



BEST MODELLING

INPUT

```
47
48  # 5. Menentukan model terbaik (berdasarkan R^2 terbesar dan RMSE terkecil)
49  best_model <- metrics_df[which.max(metrics_df$R2), "Model"]
50  print(paste("Model terbaik adalah:", best_model))
51
```

BEST MODELLING OUTPUT

values	
best_model	"Model 1"

PREDICTION MODEL

INPUT

```
51
52  # 6. Visualisasi prediksi model terbaik
53  best_model_object <- switch(
54    best_model,
55    "Model 1" = model1,
56    "Model 2" = model2,
57    "Model 3" = model3
58  )
59
```

METRICS 3

PREDICTION MODEL OUTPUT

best_model_object	list [12] (S3: lm)	List of length 12
coefficients	double [3]	257.219 -0.448 -0.724
residuals	double [12002]	-250 -251 -250 -248 -249 -249 ...
effects	double [12002]	-27488.6 -64.6 -91.4 -250.6 -249.6 -249.9 ...
rank	integer [1]	3
fitted.values	double [12002]	251 252 251 249 250 250 ...
assign	integer [3]	0 1 2
qr	list [5] (S3: qr)	List of length 5
df.residual	integer [1]	11999
xlevels	list [0]	List of length 0
call	language	lm(formula = student_id ~ stress_level + sleep_hours, data = trainData)
terms	formula	student_id ~ stress_level + sleep_hours
model	list [12002 x 3] (S3: data.frame)	A data.frame with 12002 rows and 3 columns

THANK YOU

More Information : Github Repository