# Student Score Prediction

Presented by [Rana Awad Abdel Aziz Saad]

# Agenda

# Project Overview

Goal: Predict student scores based on study hours.
Tools Used: Python, Pandas, NumPy, Matplotlib, Scikit-learn
Algorithm: Linear Regression

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

from sklearn.metrics import mean_squared_error, r2_score
```
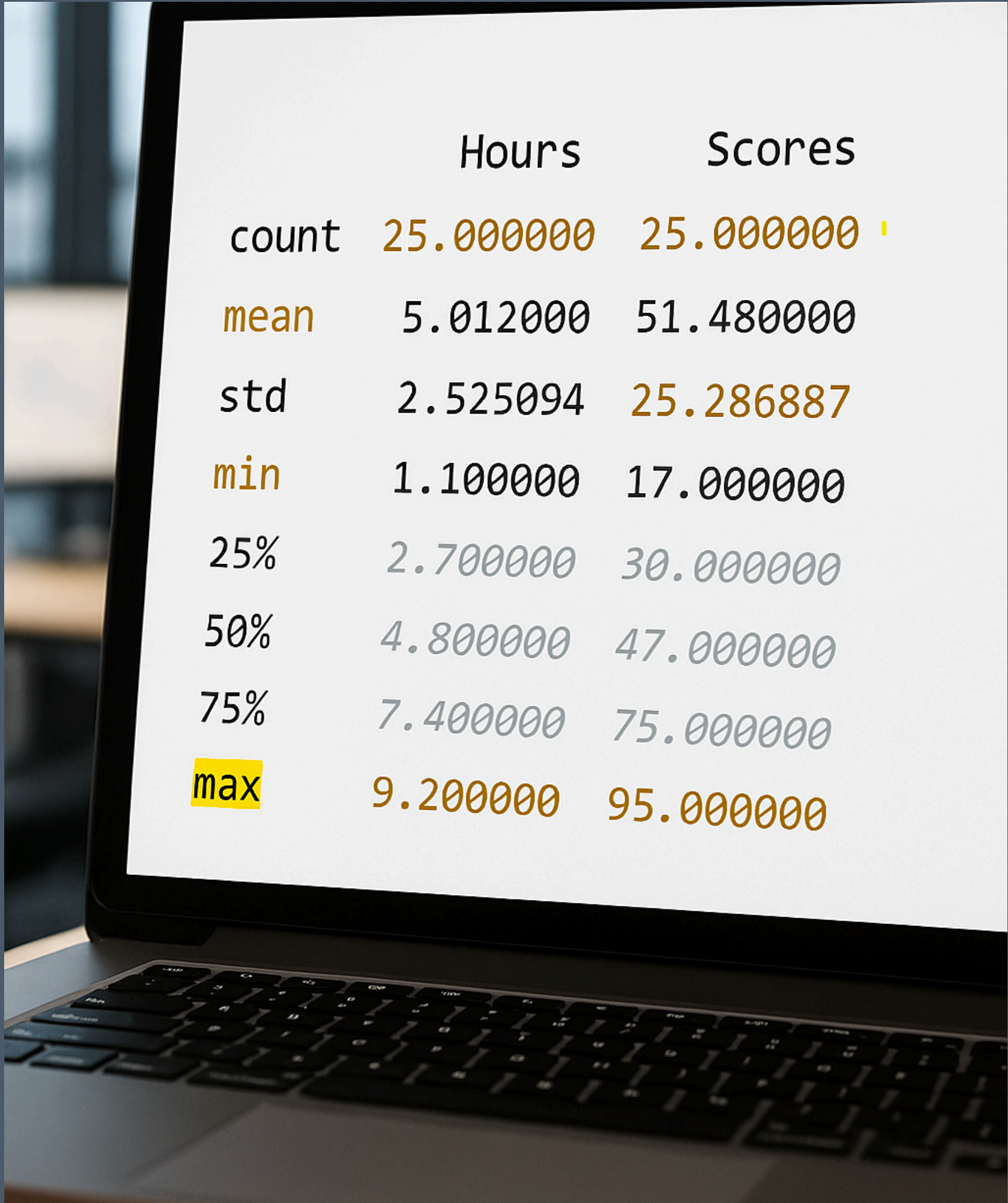
# Overview of the dataset used

The dataset for the Student Score Prediction Project consists of **25 records** detailing the relationship between study hours and exam scores. Each record contains two columns: one representing **study hours** and the other representing **exam scores**. Importantly, the dataset has no missing values, ensuring a complete analysis.

The mean study hours recorded is approximately **5 hours**, while the mean exam score is around **51 points**. This average indicates a moderate level of study time correlated with exam performance. The dataset provides a solid foundation for applying linear regression techniques, allowing us to explore how variations in study hours can predict student performance on exams. The next step involves preprocessing this data to visualize its characteristics effectively.

|       | 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 |

# Data Processing

- We checked for missing values – none were found

```
data.isnull().sum()
```

- we split the data into:
- Training Set (80%) to train the model
- Testing Set (20%) to evaluate its accuracy

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

# Model Training

Predicted Score = 9.78 × Hours + 2.48

We used the Linear Regression model which
- learns the relationship between hours (input) and scores (output).
- finds a straight line (y = a*x + b) that best fits the data.

Here,

x = Hours
y^ = predicted score
a = slope (coef)
b = intercept

```python
model = LinearRegression()
model.fit(X_train, y_train^)
```
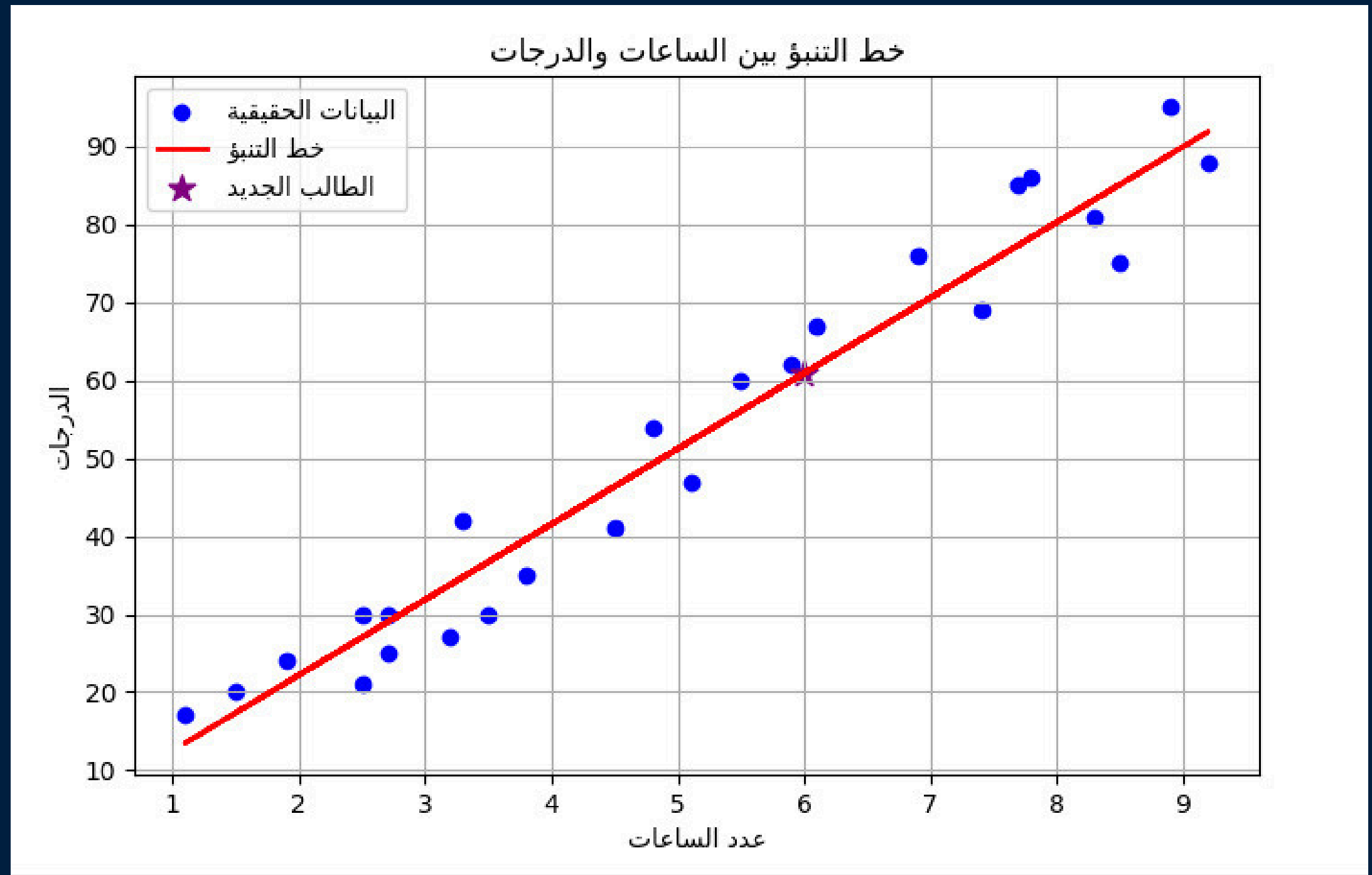
```python
print("coef =", model.coef_[0])
print("intercept =", model.intercept_)
```

# Data Visualization of Scores

## Hours vs. Scores Scatter Plot

This plot shows a **strong positive correlation**.

```
plt.scatter(data['Hours'], data['Scores'])
```



خط التنبؤ بين الساعات والدرجات

Legend:
- البيانات الحقيقية
- خط التنبؤ
- الطالب الجديد

Y-axis: الدرجات
X-axis: عدد الساعات

Blue = Real data
Red = Prediction line
Purple = New student entered manually

# Model Evaluation

- We used two metrics:
- MSE (Mean Squared Error) = 18.94 →
- $R^2$ (R-squared) = 0.967 → means the model explains 96.7% of the relationship between hours and scores.

**18.94**

Mean Squared Error (MSE)
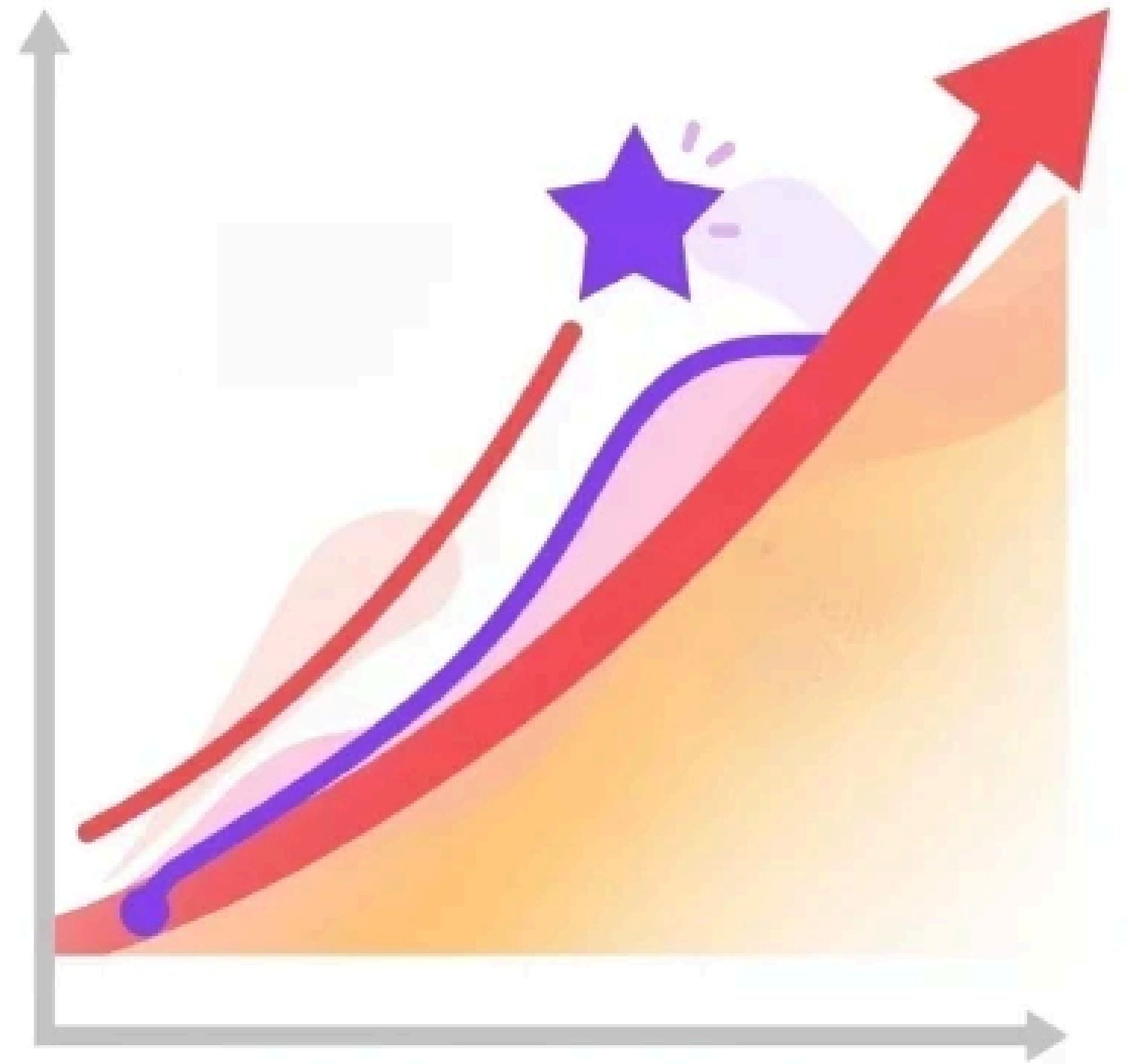measures average prediction error
lower is the better

**0.96**

R-squared Value
the model explains 96.7% of the
relationship between hours and
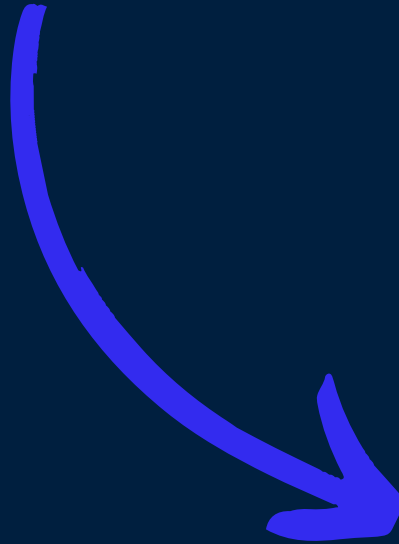scores. Closer than 1 is the better

# Prediction

- Input = 6 study hours "manually "
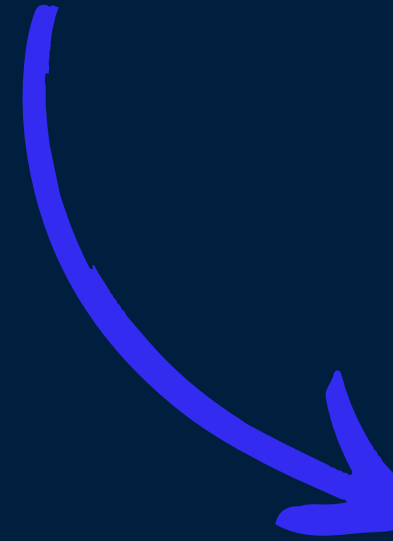- Predicted score = 9.78×6 + 2.48 ≈ 60.92

# Challenges

- Small dataset (only 25 records).
- Simple model (only one feature).
- Real-life data may include other factors like attendance, sleep, or stress.

# Conclusion

This project demonstrates the power of Machine Learning in education

It provides valuable insights into how study time affects student performance.
More study hours → Higher score

# Thank You for Your Attention

**Email**

ranaawad20199@gmail.com

**Github**