รายงานผลการทดลอง (Results)

1. การเตรียมข้อมูล (Preparing Data)

ในการทดลองครั้งนี้ใช้ชุดข้อมูล **data.csv** ซึ่งประกอบด้วยตัวแปรทั้งหมด 4 ตัว ได้แก่

- Duration
- Pulse
- Maxpulse
- Calories

จากการตรวจสอบข้อมูลพบว่า คอลัมน์ **Calories** มีข้อมูลสูญหาย (Missing Values) จำนวน **5 แถว** จึงได้ทำการลบแถวที่มีค่า NaN ออกก่อนนำข้อมูลไปสร้างโมเดล

```
df = df.dropna(subset=["Calories"])
```

หลังจากลบข้อมูลที่ไม่สมบูรณ์แล้ว เหลือข้อมูลทั้งหมด **164 แถว**

2. Simple Linear Regression

ได้สร้างโมเดล Simple Linear Regression โดยใช้

- ตัวแปรอิสระ (Feature) คือ **Duration**
- ตัวแปรตาม (Target) คือ **Calories**

ผลการทดลองแสดงให้เห็นว่า
โมเดลสามารถสร้างเส้นตรงที่อธิบายความสัมพันธ์ระหว่างระยะเวลาการออกกำลังกายกับปริมาณแคลอรีที่เผาผลาญได้ โดยใช้สมการเชิงเส้น

$Calories = b_0 + b_1(Duration)$Calories=b0 +b1 (Duration)

และแสดงผลเป็นกราฟ Regression Line เปรียบเทียบกับข้อมูลจริง

3. Multiple Linear Regression

ได้สร้างโมเดล Multiple Linear Regression โดยใช้

- Duration
- Pulse
- Maxpulse

เป็นตัวแปรอิสระ

และใช้

- Calories

เป็นตัวแปรตาม

ผลการทดลองพบว่า การใช้ตัวแปรหลายตัวช่วยให้โมเดลสามารถอธิบายความสัมพันธ์ของข้อมูลได้ดีกว่า Simple Linear Regression เนื่องจากการเผาผลาญแคลอรีขึ้นอยู่กับหลายปัจจัย ไม่ใช่เฉพาะระยะเวลาในการออกกำลังกายเพียงอย่างเดียว

4. Logistic Regression (Classification)

ในการทดลอง Classification ได้แปลงข้อมูล Calories ให้เป็น 2 กลุ่ม ได้แก่

- High Calories (Calories ≥ 300)
- Low Calories (Calories < 300)

จากนั้นใช้

- Duration
- Pulse
- Maxpulse

เป็นตัวแปรอิสระในการสร้างโมเดล Logistic Regression

ผลการทดลองสามารถจำแนกข้อมูลออกเป็น 2 กลุ่มได้ และสามารถประเมินผลด้วย Accuracy, Confusion Matrix และ Classification Report

5. Confusion Matrix

ผลการทดลองแสดง Confusion Matrix ซึ่งใช้วัดจำนวนข้อมูลที่โมเดลจำแนกได้ถูกต้องและผิดพลาด

ประกอบด้วย

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

โดยค่าที่อยู่บนแนวทแยงของตารางแสดงถึงจำนวนข้อมูลที่โมเดลทำนายได้ถูกต้อง ส่วนค่าที่อยู่นอกแนวทแยงคือข้อมูลที่โมเดลทำนายผิด

6. Model Performance Metrics

สำหรับ Regression ได้ประเมินผลด้วย

- R^2 Score
- MAE
- MSE
- RMSE

โดย

- ค่า R^2 ยิ่งมีค่าใกล้ 1 แสดงว่าโมเดลสามารถอธิบายข้อมูลได้ดี
- ค่า MAE, MSE และ RMSE ยิ่งมีค่าน้อยยิ่งแสดงว่าโมเดลมีความคลาดเคลื่อนต่ำ

สำหรับ Classification ได้ประเมินผลด้วย

- Accuracy
- Precision
- Recall
- F1-score

รวมถึง Confusion Matrix เพื่อวิเคราะห์ประสิทธิภาพของโมเดลในการจำแนกข้อมูล

สรุปผลการทดลอง (Conclusion)

จากการทดลองพบว่า ชุดข้อมูล **data.csv** สามารถนำมาสร้างโมเดล Machine Learning ได้ทั้งแบบ Regression และ Classification โดยก่อนการสร้างโมเดลจำเป็นต้องจัดการข้อมูลที่หายไป (Missing Values) เพื่อให้การฝึกโมเดลมีประสิทธิภาพ

สำหรับ Regression พบว่า Multiple Linear Regression ให้ผลการทำนายดีกว่า Simple Linear Regression เนื่องจากใช้ข้อมูลหลายตัวแปรร่วมกันในการพยากรณ์ค่า Calories ส่วน Logistic Regression สามารถนำมาใช้จำแนกข้อมูลออกเป็นกลุ่ม High Calories และ Low Calories ได้อย่างมีประสิทธิภาพ โดยใช้ตัวชี้วัด Accuracy, Precision, Recall, F1-score และ Confusion Matrix ในการประเมินผล

โดยรวมแล้ว การเลือกใช้ Regression หรือ Classification ขึ้นอยู่กับลักษณะของตัวแปรเป้าหมาย หากต้องการทำนายค่าตัวเลขควรใช้ Regression แต่หากต้องการจำแนกข้อมูลออกเป็นประเภท ควรใช้ Classification ครับ