รายงานสรุปผลการทดลอง LAB-04

Machine Learning Classification & Clustering (FPL Dataset)

1. วัตถุประสงค์และภาพรวมการทดลอง

การทดลองนี้มีวัตถุประสงค์เพื่อประยุกต์ใช้อัลกอริทึม **k-Nearest Neighbors (KNN)** ในการจำแนกประเภท (Classification) และ **K-Means Clustering** ในการจัดกลุ่มสไตล์การเล่น (Clustering) บนชุดข้อมูลสถิตินักเตะ **Fantasy Premier League (FPL)** โดยเน้นการเปรียบเทียบประสิทธิภาพของโมเดล ณ ค่า *k* ที่แตกต่างกัน และวิเคราะห์ขอบเขตพฤติกรรมข้อมูลของผู้เล่นตามสถิติการเล่นจริง

2. ข้อมูลและการเตรียมประมวลผล (Dataset & Preprocessing)

สรุปข้อมูลหลัก

- **ชุดข้อมูล:** `fpl.csv` (จำนวนผู้เล่นทั้งหมด 567 คน)
- **การกรองข้อมูล (Data Cleaning):** คัดเลือกเฉพาะผู้เล่นที่มีเวลาลงสนาม (`minutes`) ≥ 180 นาที เพื่อตัดข้อผิดพลาดจากผู้เล่นสำรองที่ไม่มีสถิติ ส่งผลให้เหลือข้อมูลคุณภาพจำนวน **340 คน**
- **การแบ่งสัดส่วนชุดข้อมูล:** แบ่งชุดฝึกฝน (Train set 70% = 238 คน) และชุดทดสอบ (Test set 30% = 102 คน) ด้วยวิธี Stratified Sampling
- **การทำสเกลข้อมูล (Feature Scaling):** ใช้ `StandardScaler` ปรับข้อมูล 14 คุณลักษณะ (เช่น `goals_scored`, `assists`, `tackles`, `recoveries`, `ict_index`) ให้อยู่ในรูป *Z*-score ($\mu = 0$, $\sigma = 1$) เพื่อป้องกันไม่ให้ฟีเจอร์ที่มีสเกลใหญ่ข่มระยะทางในการคำนวณ Euclidean Distance

สัดส่วนตำแหน่งนักเตะในชุดข้อมูล (Target Classes)

MID (กองกลาง): 155 คน (45.6%)	DEF (กองหลัง): 118 คน (34.7%)	FWD (กองหน้า): 39 คน (11.5%)	GKP (ผู้รักษาประตู): 28 คน (8.2%)

3. สรุปผลการทดลองการจำแนกประเภท (KNN Classification)

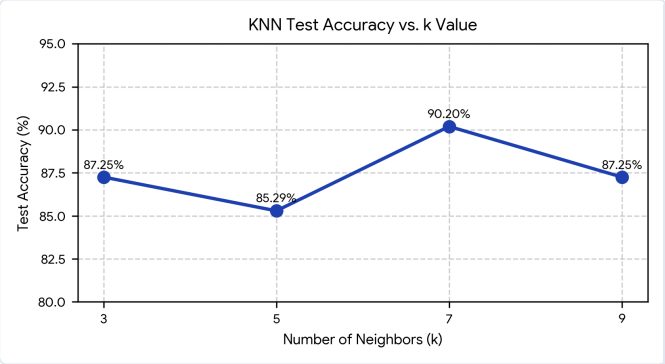
ตารางเปรียบเทียบประสิทธิภาพตามค่า *k* (Test Set = 102 คน)

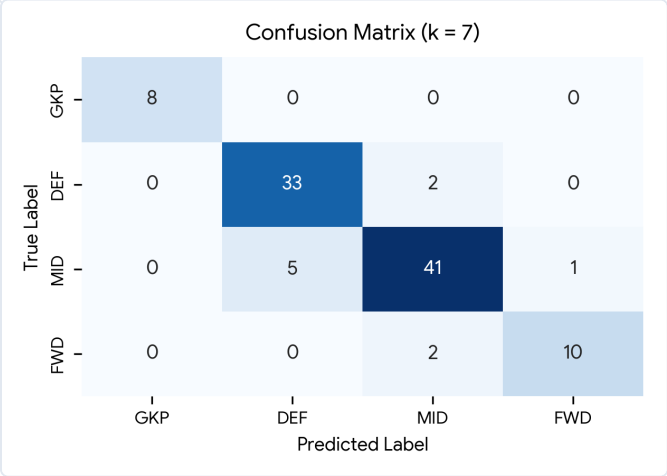
ค่า *k* (n_neighbors)	Test Accuracy Score	Accuracy (%)	จำนวนเคสที่ทายถูกต้อง	จำนวนเคสที่ทายผิด
k = 3	0.8725	87.25%	89 / 102	13 คน
k = 5	0.8529	85.29%	87 / 102	15 คน
***k* = 7 (Best)**	**0.9020**	**90.20%**	**92 / 102**	**10 คน**
k = 9	0.8725	87.25%	89 / 102	13 คน

✔ ผลลัพธ์ที่ดีที่สุด: โมเดล ณ *k* = 7 ให้ความถูกต้องสูงสุดที่ 90.20% (ทำนายถูกต้อง 92 จาก 102 เคส)

รายละเอียดการประเมินผลแยกตามตำแหน่ง (Classification Report at *k* = 7)

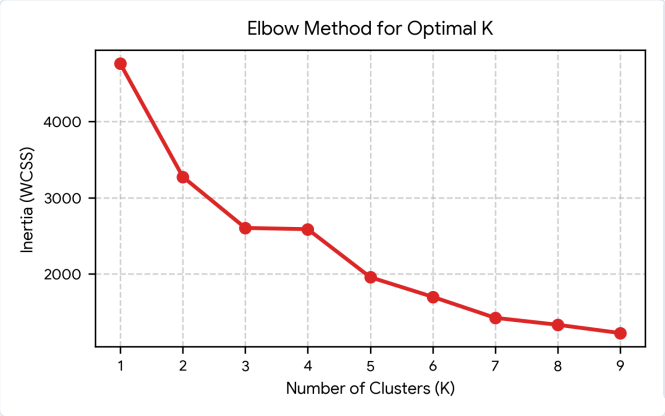
ตำแหน่ง (Position)	Precision	Recall	F1-Score	จำนวนทดสอบ (Support)
GKP (ผู้รักษาประตู)	1.00	1.00	1.00	8
DEF (กองหลัง)	0.87	0.94	0.90	35
MID (กองกลาง)	0.91	0.87	0.89	47
FWD (กองหน้า)	0.91	0.83	0.87	12
ค่าเฉลี่ยภาพรวม (Accuracy)	**-**	**-**	**0.90**	**102**

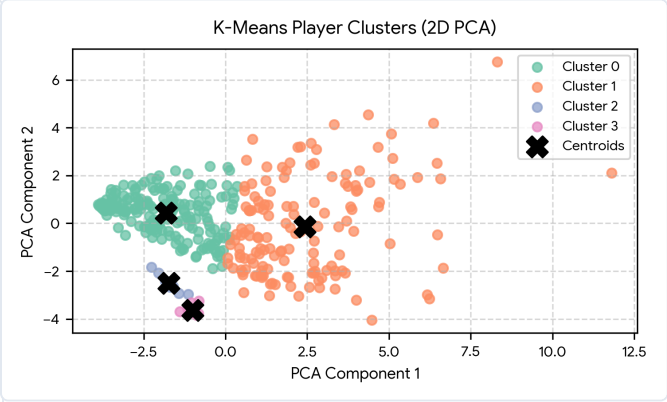




4. สรุปผลการจัดกลุ่ม (K-Means Clustering)

- **การหาจำนวนกลุ่มที่เหมาะสม (Elbow Method):** ค่า Inertia (Within-Cluster Sum of Squares) ลดลงอย่างต่อเนื่องและเกิดจุดหักเหที่เด่นชัด ณ $K = 4$ ซึ่งสอดคล้องกับตำแหน่งหลักทั้ง 4 ในกีฬาฟุตบอล
- **การลดมิติข้อมูลเพื่อแสดงผล (PCA Visualization):** เมื่อฉายข้อมูล 14 มิติลงบนพื้นที่ 2 มิติผ่าน Principal Component Analysis (PCA) พบว่า:
  - กลุ่มผู้รักษาประตู (GKP) แยกตัวออกจากกลุ่มอื่นอย่างเด็ดขาด
  - กลุ่มกองหลัง (DEF) และกองกลาง (MID) มีพื้นที่ทับซ้อนกันบางส่วน เนื่องจากสถิติการตัดบอลและการทำเกมของกองหลังวิงแบ็กและกองกลางตัวรับมีความใกล้เคียงกัน





5. อภิปรายและสรุปผลการทดลอง (Discussion & Conclusion)

- **อิทธิพลของค่า *k* ใน KNN:**
  - ค่า k ที่น้อยเกินไป ($k = 3, 5$) ทำให้โมเดลไวต่อ Noise หรือผู้เล่นที่มีสถิติสุดโต่งเฉพาะบุคคล
  - ค่า $k = 7$ เป็นระดับที่สร้างขอบเขตการตัดสินใจ (Decision Boundary) ได้เรียบและสมดุลที่สุด ส่งผลให้ได้ค่าความถูกต้องสูงสุด 90.20%
- **คุณลักษณะของฟีเจอร์เฉพาะตำแหน่ง:**
  - สถิติเฉพาะทาง เช่น `saves` ช่วยให้จำแนกผู้รักษาประตูได้ถูกต้อง 100%
  - ความผิดพลาดส่วนใหญ่เกิดขึ้นระหว่าง DEF และ MID เนื่องจากการเปลี่ยนผ่านบทบาทในฟุตบอลสมัยใหม่ที่กองหลังมีส่วนร่วมกับการเกมรุก และกองกลางมีบทบาทเกมรับมากขึ้น
- **การจัดโครงสร้างโปรเจกต์และการเวอร์ชันซอฟต์แวร์:**
  - การแยกโมดูล `classification` และ `clustering` ออกเป็นโครงสร้างระบบระเบียบ ช่วยให้ง่ายต่อการซ่อมบำรุงและรันประมวลผลซ้ำผ่านระบบติดตามเวอร์ชัน Git/GitHub