รายงานการสำรวจข้อมูล

ผลกระทบของ Generative AI
ต่อผลการเรียนและความเป็นอยู่ที่ดีของนักเรียน

(AI Student Impact Dataset — LAB 1: Dataset Exploration)

ชุดข้อมูล: ai_student_impact_dataset.csv | ขนาด: 50,000 แถว x 16 คอลัมน์ |
จัดทำโดย: TARo

1. บทนำ

ชุดข้อมูลนี้รวบรวมพฤติกรรมการใช้เครื่องมือ Generative AI (เช่น ChatGPT, Copilot, Gemini) ของนักเรียนจำนวน 50,000 คน ครอบคลุมทั้งด้านผลการเรียน (GPA ก่อน-หลังภาคการศึกษา) พฤติกรรมการใช้ AI พฤติกรรมการเรียนแบบดั้งเดิม บริบทเชิงนโยบายของสถาบัน และสุขภาพจิต/ความเป็นอยู่ที่ดี เช่น ความวิตกกังวลในการสอบและความเสี่ยงต่อภาวะหมดไฟ (Burnout) รายงานฉบับนี้สรุปผลจากขั้นตอนการสำรวจข้อมูลเบื้องต้น (LAB 1: Dataset Exploration)

2. โครงสร้างของชุดข้อมูล

ชุดข้อมูลมีขนาด 50,000 แถว x 16 คอลัมน์ เป็นชุดข้อมูลที่สมบูรณ์ กล่าวคือ ไม่พบค่าที่ขาดหายไป (missing values = 0) และ ไม่พบข้อมูลที่ซ้ำกัน (duplicate records = 0) ตลอดทั้งชุดข้อมูล จึงไม่จำเป็นต้องทำความสะอาดข้อมูลขั้นพื้นฐาน ก่อนนำไปวิเคราะห์เพิ่มเติม

2.1 กลุ่มคุณลักษณะ (Feature Groups)

กลุ่ม	คอลัมน์ตัวอย่าง	ชนิดข้อมูล
ตัวระบุ	Student_ID	จำนวนเต็ม
ประวัติทางวิชาการ	Major_Category, Year_of_Study, Pre/Post_Semester_GPA	หมวดหมู่ / ทศนิยม

พฤติกรรมการใช้ AI	Weekly_GenAI_Hours, Primary_Use_Case, Prompt_Engineering_Skill, Tool_Diversity	ทศนิยม / หมวดหมู่ / จำนวนเต็ม
พฤติกรรมการเรียน	Traditional_Study_Hours, Perceived_AI_Dependency	ทศนิยม / จำนวนเต็ม
บริบทของสถาบัน	Institutional_Policy	หมวดหมู่
สุขภาพจิต/ความเป็นอยู่ที่ดี	Anxiety_Level_During_Exams, Skill_Retention_Score, Burnout_Risk_Level	จำนวนเต็ม / ทศนิยม / หมวดหมู่

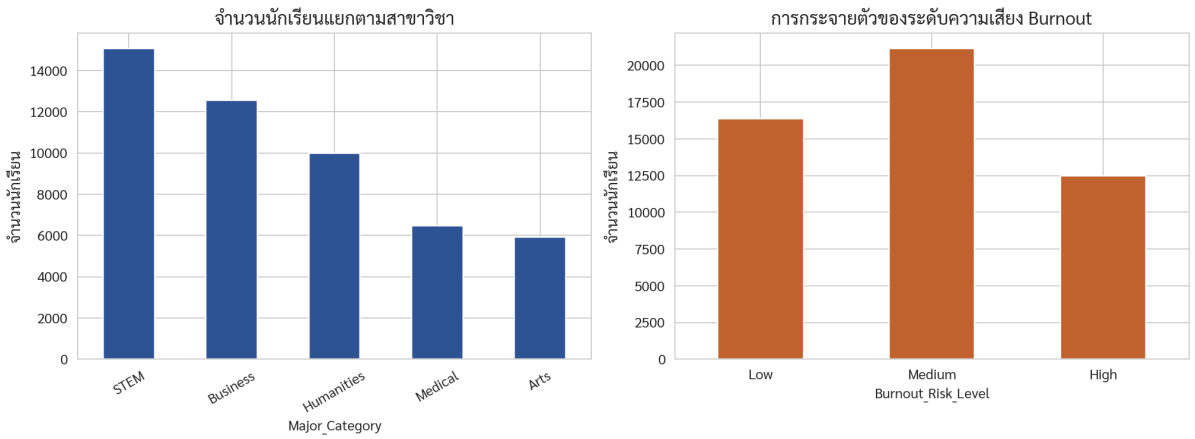
3. สถิติเชิงสรุปของตัวแปรตัวเลขหลัก

ตัวแปร	ต่ำสุด	เฉลี่ย	สูงสุด
Pre_Semester_GPA	1.18	3.15	4.00
Post_Semester_GPA	1.00	3.35	4.00
Weekly_GenAI_Hours (ชม./สัปดาห์)	0.00	8.43	40.00
Traditional_Study_Hours (ชม./สัปดาห์)	1.00	11.21	35.86
Skill_Retention_Score (0-100)	10.78	75.80	100.00
Perceived_AI_Dependency (1-10)	1	3.51	10
Anxiety_Level_During_Exams (1-10)	1	4.27	10

4. การกระจายตัวของตัวแปรหมวดหมู่ (Class Distribution)

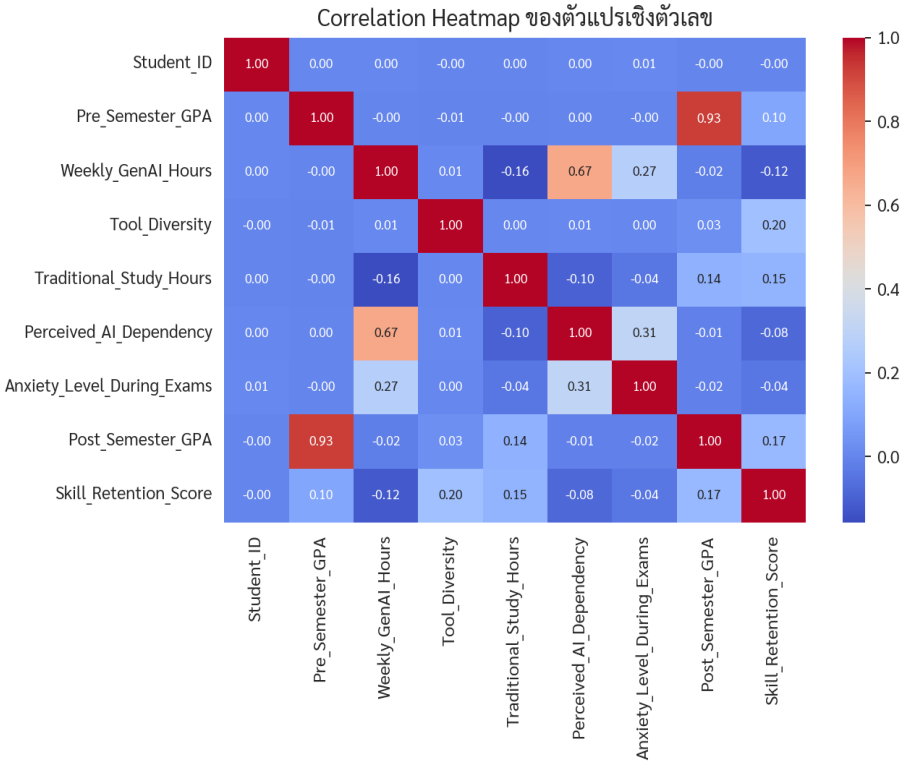
สาขาวิชาที่มีจำนวนนักเรียนมากที่สุดคือ STEM (15,059 คน) รองลงมาคือ Business (12,538 คน), Humanities (9,994 คน), Medical (6,476 คน) และ Arts (5,933 คน) ตามลำดับ ในด้านความเสี่ยงต่อภาวะหมดไฟ นักเรียนส่วนใหญ่จัดอยู่ในระดับ Medium (21,144 คน) ตามด้วย Low (16,369 คน) และ High (12,487 คน) ขณะที่ด้านนโยบายของสถาบัน

สถาบันส่วนใหญ่ อนุญาตให้ใช้ AI โดยต้องอ้างอิงทีมา (Allowed_With_Citation, 25,224 คน) มากกว่าการส่งเสริมอย่างเต็มที (14,988 คน) หรือการห้ามใช้โดยเด็ดขาด (9,788 คน)



ภาพที 1: การกระจายตัวของนักเรียนตามสาขาวิชา และระดับความเสี่ยง Burnout

5. ความสัมพันธ์ระหว่างตัวแปรเชิงตัวเลข



ภาพที่ 2: Correlation Heatmap ของตัวแปรเชิงตัวเลขทั้งหมด

จากค่าสัมประสิทธิ์สหสัมพันธ์ พบว่า Post_Semester_GPA มีความสัมพันธ์เชิงบวกสูงมากกับ Pre_Semester_GPA ($r \approx 0.93$) ซึ่งสมเหตุสมผลเนื่องจากพื้นฐานผลการเรียนเดิมมักส่งผลต่อผลการเรียนถัดไป ในทางกลับกัน Weekly_GenAI_Hours มีความสัมพันธ์กับ Post_Semester_GPA ที่ค่อนข้างต่ำ ($r \approx -0.02$) เช่นเดียวกับ Perceived_AI_Dependency ($r \approx -0.01$) และ Anxiety_Level_During_Exams ($r \approx -0.02$) ซึ่งบ่งชี้ว่าจากข้อมูลชุดนี้ จำนวนชั่วโมงการใช้ AI เพียงอย่างเดียวไม่ได้มีความสัมพันธ์เชิงเส้นที่ชัดเจนกับ ผลการเรียนปลายภาค และควรพิจารณาปัจจัยร่วมอื่น ๆ เพิ่มเติมในการวิเคราะห์เชิงลึกต่อไป

6. สรุปผลและข้อเสนอแนะสำหรับการวิเคราะห์ขั้นถัดไป

- ชุดข้อมูลมีความสมบูรณ์สูง (ไม่มีค่าว่าง/ไม่มีข้อมูลซ้ำ) จึงพร้อมสำหรับขั้นตอน Feature Engineering และการสร้างแบบจำลอง (modeling) ได้ทันที
- ควรตรวจสอบขอบเขตค่าที่สมเหตุสมผล (valid range) ของแต่ละคอลัมน์ เช่น GPA ควรอยู่ระหว่าง 0.00-4.00 ก่อนนำไปใช้งานจริง เพื่อป้องกันข้อมูลผิดปกติที่อาจหลุดรอดจากการตรวจสอบเบื้องต้น
- ตัวแปร Burnout_Risk_Level และ Prompt_Engineering_Skill เหมาะสำหรับ Label Encoding เนื่องจากมีลำดับ (ordinal) ขณะที่ Major_Category และ

Institutional_Policy เหมาะสำหรับ One-Hot Encoding เนื่องจากไม่มีลำดับ (nominal)

- ควรวิเคราะห์เพิ่มเติมในเชิงกลุ่มย่อย (subgroup analysis) เช่น
เปรียบเทียบ Post-Pre GPA แยกตามสาขาวิชา
หรือแยกตามนโยบายของสถาบัน
เพื่อหาปัจจัยที่ส่งผลต่อผลการเรียนได้ชัดเจนยิ่งขึ้น