Final Project

# การศึกษาปัจจัยที่มีผลต่อความเครียดของนักศึกษาวิศวกรรมคอมพิวเตอร์ มจธ.

# Identifying factors affecting stress levels among Computer Engineering students at KMUTT


คณะผู้จัดทำ

| | | |
|---|---|---|
| นางสาวกัญญวรรณ | ทยานิพันธ์ | 66070501004 |
| นางสาวกัลยา | สุทธปัญญา | 66070501005 |
| นางสาวจิรพรรณ | ศักดิ์พิสุทธิพงศ์ | 66070501009 |
| นางสาวเปมิกา | จงขวัญยืน | 66070501034 |
| นางสาวศิรประภา | ชินวงษ์ทัน | 66070501055 |
| นายกันต์ดนัย | ศรีวัฒนะ | 65070507203 |
| นายญาณภัทร | พึ่งแสงธรรมเลิศ | 65070507231 |

เสนอ

ดร. สัญญ์สิริ ธารประดับ


ภาคเรียนที่ 2 ปีการศึกษา 2567

รายวิชา CPE 232 แบบจำลองข้อมูล (Data Models)


ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

# สารบัญ


**ชื่อเรื่อง** **หน้า**

**Introduction** ..... 1

วัตถุประสงค์ ..... 1

รายละเอียดการเก็บข้อมูลโดย Google Form ..... 2

**Preprocessing** ..... 4

1. Import Libraries ที่จำเป็น ..... 4

2. Import ข้อมูล ..... 4

3. ตรวจสอบภาพรวมของข้อมูล ..... 4

4. ลบและแปลงคอลัมน์ที่ไม่จำเป็น ..... 4

5. แปลงข้อความที่เป็นช่วงเวลา หรือคำอธิบายให้เป็นตัวเลข ..... 4

6. แปลงข้อมูลให้อยู่ในรูปแบบตัวเลขทศนิยม (Float) ..... 5

7. แปลงข้อมูลจากข้อความลำดับ (Ordinal) ให้เป็นตัวเลข (Ordinal Encoding) ..... 5

8. แปลงข้อมูลประเภท Binary และการจัดลำดับอื่น ๆ ..... 6

9. จัดการกับ Missing Data (ค่าที่ขาดหาย) ..... 6

10. ตรวจสอบภาพรวมของข้อมูลอีกครั้ง ..... 6

สรุปขั้นตอนการ preprocessing ..... 7

**Exploratory Data Analysis (EDA)** ..... 8

1. Import Libraries ..... 8

2. Load Dataset ..... 8

3. ตรวจสอบและปรับปรุงข้อมูล ..... 8

4. วิเคราะห์ข้อมูล ..... 8

5. สรุปข้อมูลที่ได้จากการ EDA ..... 9

6. ส่งต่อข้อมูล ..... 12

สรุปกระบวนการ EDA ..... 12

**Method** ..... 13

Model Selection ..... 13

Training & Evaluation ..... 13

การวิเคราะห์ผล (Result Analysis) ..... 16

**Result & Discussion** ..... 17

ผลการศึกษา (Results) ..... 17

การอภิปรายผล (Discussion) ..... 19

**Conclusion** ..... 25

**ภาคผนวก ก** ..... ก

**ภาคผนวก ข** ..... ข

# Introduction

ปัญหาความเครียดในหมู่นักศึกษาระดับมหาวิทยาลัยถือเป็นประเด็นสำคัญที่ส่งผลต่อสุขภาพจิต สมรรถภาพทางการศึกษา และคุณภาพชีวิตโดยรวม โดยเฉพาะในสาขาวิชาที่มีการแข่งขันสูงและความกดดันจากภาระงานวิชาการ เช่น สาขาวิศวกรรมคอมพิวเตอร์ นักศึกษามักเผชิญกับความกดดันจากงานที่ได้รับมอบหมาย โครงการกลุ่ม ความคาดหวังทางวิชาการ ตลอดจนการปรับตัวต่อเทคโนโลยีที่เปลี่ยนแปลงอย่างรวดเร็ว ซึ่งอาจนำไปสู่ความเครียดสะสม

งานศึกษานี้จัดทำขึ้นภายใต้รายวิชา CPE232 DATA MODELS ภาคการศึกษา 2/2567 เพื่อวิเคราะห์และระบุปัจจัยที่มีผลต่อระดับความเครียดของนักศึกษาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี (มจธ.) โดยใช้ข้อมูลจากแบบสอบถาม Google Form และนำไปประมวลผลในเชิงสถิติและการวิเคราะห์ข้อมูล เพื่อหาความสัมพันธ์ระหว่างตัวแปรต่าง ๆ กับระดับความเครียดของนักศึกษาและนำไปสู่การวิเคราะห์และสร้างแบบจำลองข้อมูลที่มีประสิทธิภาพอีกทั้งยังสามารถนำไปใช้ในการพัฒนามาตรการสนับสนุนนักศึกษาในอนาคต

**วัตถุประสงค์**

1. เพื่อระบุปัจจัยที่มีผลต่อระดับความเครียดของนักศึกษาวิศวกรรมคอมพิวเตอร์ มจธ. ในมิติต่าง ๆ ได้แก่ ปัจจัยด้านประชากรศาสตร์ การศึกษา พฤติกรรมการใช้ชีวิต และสุขภาพจิตและสังคม
2. เพื่อเก็บรวบรวมข้อมูลที่ครอบคลุมและเชื่อถือได้โดยใช้แบบสอบถาม Google Form ซึ่งออกแบบมาเพื่อประเมินความเครียดผ่านเครื่องมือมาตรฐาน ST-5 และตัวแปรอื่น ๆ ที่เกี่ยวข้อง
3. เพื่อนำข้อมูลที่ได้ไปใช้ในการสร้างแบบจำลองข้อมูลที่สามารถวิเคราะห์ความสัมพันธ์ระหว่างปัจจัยต่าง ๆ กับระดับความเครียด เพื่อเป็นพื้นฐานสำหรับงานวิจัยและการพัฒนามาตรการสนับสนุนนักศึกษา

**รายละเอียดการเก็บข้อมูลโดย Google Form**

การเก็บข้อมูลสำหรับการศึกษานี้ใช้ Google Form เป็นเครื่องมือหลัก โดยแบบสอบถามได้รับการออกแบบอย่างรอบคอบเพื่อให้ครอบคลุมข้อมูลที่จำเป็นสำหรับการวิเคราะห์ปัจจัยที่มีผลต่อความเครียดของนักศึกษาวิศวกรรมคอมพิวเตอร์ มจธ. จำนวนผู้ตอบแบบสอบถามทั้งหมด 102 คน ซึ่งระยะเวลาการเก็บข้อมูลอยู่ระหว่างวันที่ 19 เมษายน พ.ศ. 2568 และ วันที่ 3 พฤษภาคม พ.ศ. 2568 รายละเอียดของการเก็บข้อมูลมีดังนี้

- โครงสร้างของแบบสอบถาม (ดูภาคผนวก ก)

  แบบสอบถามประกอบด้วย 5 ส่วน รวมทั้งสิ้น 26 คำถาม ได้แก่

  - ส่วนที่ 1: ข้อมูลด้านประชากรศาสตร์รวบรวมข้อมูลพื้นฐาน เช่น รหัสนักศึกษา ซึ่งบ่งบอกระดับชั้นปี และหลักสูตรที่กำลังศึกษา และเพศเพื่อวิเคราะห์ความแตกต่างของระดับความเครียดในกลุ่มประชากรที่หลากหลาย
  - ส่วนที่ 2: ปัจจัยทางการศึกษาครอบคลุมข้อมูลเกี่ยวกับภาระงานวิชาการ ความยากของวิชาเรียน และความคาดหวังในผลการเรียน ซึ่งอาจเป็นปัจจัยสำคัญที่ก่อให้เกิดความเครียด
  - ส่วนที่ 3: พฤติกรรมการใช้ชีวิตสำรวจพฤติกรรม เช่น การนอน การออกกำลังกาย การทานอาหาร และการใช้สื่อสังคมออนไลน์ ซึ่งอาจส่งผลต่อสุขภาพจิต
  - ส่วนที่ 4: สุขภาพจิตและสังคมเก็บข้อมูลเกี่ยวกับความสัมพันธ์กับเพื่อน ครอบครัว รวมถึงการเงินที่อาจก่อให้เกิดความเครียด
  - ส่วนที่ 5: ตัวแปรเป้าหมาย (Label) ใช้แบบประเมินความเครียด ST-5 ซึ่งพัฒนาโดยกรมสุขภาพจิต กระทรวงสาธารณสุข เพื่อวัดระดับความเครียดผ่านอาการต่าง ๆ เช่น การนอนไม่หลับ ความหงุดหงิด สมาธิสั้น ความเบื่อหน่าย และการแยกตัวจากสังคม ผลจากส่วนนี้จะถูกใช้เป็นตัวแปรเป้าหมายในการสร้างแบบจำลองข้อมูล

- วัตถุประสงค์ของการเก็บข้อมูล

ข้อมูลที่รวบรวมผ่าน Google Form จะถูกใช้เพื่อวัตถุประสงค์ทางการศึกษาในรายวิชา CPE232 DATA MODELS ภาคการศึกษา 2/2567 เท่านั้น โดยมีเป้าหมายเพื่อศึกษาความสัมพันธ์ระหว่างปัจจัยต่าง ๆ และระดับความเครียดของนักศึกษา ข้อมูลทั้งหมดจะถูกเก็บรักษาเป็นความลับและไม่มีการเปิดเผยข้อมูลส่วนบุคคล

- วิธีการเก็บข้อมูล
    - ช่วงเวลาการเก็บข้อมูล: ผู้ตอบแบบสอบถามได้รับคำแนะนำให้กรอกข้อมูลในช่วงเวลาปกติที่ไม่ใช่ช่วงสอบ เพื่อให้ได้ข้อมูลที่สะท้อนถึงสถานการณ์ทั่วไปและหลีกเลี่ยงความคลาดเคลื่อนจากความเครียดในช่วงสอบ
    - การรักษาความลับ: ข้อมูลส่วนบุคคลทั้งหมดจะถูกเก็บรักษาด้วยความปลอดภัยและใช้เพื่อวัตถุประสงค์ทางวิชาการเท่านั้น
    - เครื่องมือที่ใช้: แบบประเมิน ST-5 เป็นเครื่องมือคัดกรองมาตรฐานที่เชื่อถือได้ ซึ่งช่วยให้สามารถวัดระดับความเครียดได้อย่างเป็นระบบและเป็นที่ยอมรับในระดับสากล
- การนำข้อมูลไปใช้

ข้อมูลที่รวบรวมได้จะถูกนำไปวิเคราะห์เพื่อหาความสัมพันธ์ระหว่างปัจจัยต่าง ๆ กับระดับความเครียด โดยใช้เทคนิคการวิเคราะห์ข้อมูลและการสร้างแบบจำลองในรายวิชา CPE232 ผลจากการวิเคราะห์จะช่วยให้เข้าใจถึงปัจจัยที่สำคัญที่สุดที่ส่งผลต่อความเครียดของนักศึกษา และสามารถนำไปสู่การพัฒนามาตรการหรือแนวทางในการลดความเครียดในอนาคต

## Preprocessing

เนื่องจากแบบสอบถามถูกออกแบบมาอย่างดีในเชิงโครงสร้าง โดยเฉพาะการกำหนดช่วงคำตอบที่ชัดเจนและจำกัดช่วงอย่างเหมาะสม ซึ่งช่วยให้การเตรียมข้อมูลมีความสะดวกและลดความคลาดเคลื่อนของข้อมูลที่ได้ ขั้นตอนที่ใช้ในการ preprocess ข้อมูลมีดังนี้

**1. Import Libraries ที่จำเป็น**

ในขั้นแรก import ไลบรารีที่จำเป็นในการจัดการและวิเคราะห์ข้อมูล ได้แก่ numpy ใช้จัดการกับข้อมูลเชิงตัวเลข pandas ใช้จัดการกับข้อมูลในลักษณะตาราง (DataFrame) และ sklearn.preprocessing เพื่อแปลงค่าข้อความเป็นตัวเลข

**2. Import ข้อมูล**

เราทำการอ่านไฟล์ .csv ซึ่งเป็นผลลัพธ์ที่ได้จาก Google Form มาเก็บไว้ในรูปแบบของ DataFrame ด้วยคำสั่ง pd.read_csv() โดยตั้งชื่อว่า df

```
df = pd.read_csv('final_rawdata.csv')
```

**3. ตรวจสอบภาพรวมของข้อมูล**

ตรวจสอบชนิดข้อมูลและภาพรวมของ dataset ด้วยคำสั่ง df.info() ผลลัพธ์แสดงว่า dataset ประกอบด้วย 102 แถว และ 27 คอลัมน์ ในจำนวนนี้มี 14 คอลัมน์ที่มีชนิดข้อมูลเป็น object (string) ซึ่งจำเป็นต้องถูกแปลงให้เป็นข้อมูลเชิงตัวเลขเพื่อใช้งานในโมเดล Machine Learning ต่อไป

**4. ลบและแปลงคอลัมน์ที่ไม่จำเป็น**

1. ลบคอลัมน์ที่ไม่มีความสำคัญต่อการวิเคราะห์ เช่น Timestamp เพราะไม่ได้มีผลต่อพฤติกรรมที่ต้องการวิเคราะห์
2. เพิ่มคอลัมน์ “ชั้นปี” โดยดึงข้อมูล 2 ตัวแรกจากรหัสนักศึกษา ซึ่งเป็นรหัสประจำปีที่เข้าศึกษา และ map เป็นชั้นปี (เช่น รหัสขึ้นต้นด้วย 64 หมายถึงนักศึกษาชั้นปีที่ 2 ในปี 2025)

**5. แปลงข้อความที่เป็นช่วงเวลา หรือคำอธิบายให้เป็นตัวเลข**

แบบสอบถามมีการใช้ข้อความเพื่อแทนช่วงค่าบางช่วง เช่น “มากกว่า 40 ชั่วโมง” หรือ “ทุกมื้ออาหาร” ซึ่งไม่สามารถนำไปวิเคราะห์เชิงตัวเลขได้โดยตรง เราจึงทำการแปลงค่าดังกล่าวให้เป็นตัวเลขประมาณการณ์ เช่น

- “มากกว่า 40 ชั่วโมง” ⟶ 41
- “มากกว่า 12 วิชา” ⟶ 13
- “ทุกมื้ออาหาร” ⟶ 21 (อนุมานว่าหมายถึง 3 มื้อต่อวัน × 7 วัน)

- “วันละมากกว่า 3 แก้ว” ⟶ 28 (อนุมานว่าหมายถึง 4 แก้ว × 7 วัน)
- “วันละมากกว่า 3 ชั่วโมง” ⟶ 21 (อนุมานว่า 3 ชั่วโมง × 7 วัน)

โดยมีการใช้คำสั่ง replace(), map(), to_numeric ของ pandas และ LabelEncoder ของ sklearn.preprocessing เพื่อแปลงข้อความเหล่านี้ให้เป็นค่าตัวเลขตามสมมุติฐาน

**6. แปลงข้อมูลให้อยู่ในรูปแบบตัวเลขทศนิยม (Float)**

หลังจากทำการ replace ข้อความแล้ว เราทำการแปลงคอลัมน์ที่เกี่ยวข้อง เช่น “จำนวนชั่วโมงต่อสัปดาห์ที่ใช้ทำกิจกรรมอื่นนอกเหนือจากการเรียน” หรือ “จำนวนวิชาที่ลงทะเบียนเรียนในเทอมปัจจุบัน” ให้กลายเป็นชนิด float ด้วย pd.to_numeric() เพื่อให้สามารถนำไปใช้ในโมเดล Logistic Regression ซึ่งรองรับเฉพาะข้อมูลเชิงตัวเลข

**7. แปลงข้อมูลจากข้อความลำดับ (Ordinal) ให้เป็นตัวเลข (Ordinal Encoding)**

บางคำตอบในแบบสอบถามอยู่ในรูปแบบของลำดับขั้น (ordinal scale) เช่น ความพึงพอใจ หรือการประเมินตนเอง ซึ่งสามารถแปลงเป็นตัวเลขได้ตามลำดับความสูง-ต่ำ เช่น

ตัวอย่าง:

- GPAX:
    - “มากกว่า 3.6” ⟶ 1
    - “3.1 - 3.5” ⟶ 2
    - “2.6 - 3.0” ⟶ 3
    - “2.0 - 2.5” ⟶ 4
    - “ต่ำกว่า 2.0” ⟶ 5
- ระยะเวลาการนอน:
    - “น้อยกว่า 4 ชั่วโมง” ⟶ 1
    - “4-5 ชั่วโมง” ⟶ 2
    - “6-7 ชั่วโมง” ⟶ 3
    - “มากกว่า 7 ชั่วโมง” ⟶ 4
- ความถี่ในการออกกำลังกาย:

- “ไม่ออกเลย” ⟶ 1
- “น้อยกว่า 1 ครั้งต่อสัปดาห์” ⟶ 2
- “1-2 ครั้งต่อสัปดาห์” ⟶ 3
- “3-5 ครั้งต่อสัปดาห์” ⟶ 4
- “มากกว่า 5 ครั้งต่อสัปดาห์” ⟶ 5

- การใช้โซเชียลมีเดีย:
  - “ไม่เล่นเลย” ⟶ 0
  - “วันละไม่เกิน 1 ชั่วโมง” ⟶ 1
  - “วันละ 1 - 3 ชั่วโมง” ⟶ 2
  - “วันละ 4 - 6 ชั่วโมง” ⟶ 3
  - “มากกว่า 6 ชั่วโมง” ⟶ 4

**8. แปลงข้อมูลประเภท Binary และการจัดลำดับอื่น ๆ**

1. **คอลัมน์ “เพศ”**: ใช้ LabelEncoder แปลงเป็นค่าตัวเลข เช่น 0 = ชาย, 1 = หญิง
2. **คอลัมน์ที่มีคำตอบแบบ Likert scale หรือความรู้สึก** เช่น
   a. ความรู้สึกไม่มีแรงจูงใจ
   b. ความยากในการโฟกัส
   c. ความอารมณ์แปรปรวน
   d. การมองโลกในแง่ดี/ร้าย
3. **ใช้ ordinal encoding** เช่น 1 = น้อยที่สุด ⟶ 5 = มากที่สุด

**9. จัดการกับ Missing Data (ค่าที่ขาดหาย)**

ตรวจพบว่ามีค่าหายเพียง 1 ค่าในคอลัมน์ “ระดับความเครียดในช่วง 30 วันที่ผ่านมา” ซึ่งอาจเกิดจากการที่ผู้ตอบไม่ได้กรอกข้อมูล และไม่ได้ตั้งให้เป็นคำถามบังคับใน Google Form เราจึงใช้ค่าเฉลี่ยของคอลัมน์นั้นเติมลงไปในแถวที่ขาด

**10. ตรวจสอบภาพรวมของข้อมูลอีกครั้ง**

เพื่อให้แน่ใจว่าข้อมูลพร้อมสำหรับการวิเคราะห์:

- ใช้ groupby() เพื่อดูสัดส่วนของนักศึกษาในแต่ละชั้นปี
- ใช้ describe() เพื่อดูสถิติเบื้องต้น เช่น mean, std, min, max
- แบ่งกลุ่มข้อมูลตามชั้นปีเพื่อให้สามารถนำไปวิเคราะห์เชิงเปรียบเทียบได้

**สรุปขั้นตอนการ preprocessing**

หลังจากดำเนินขั้นตอนการ preprocessing ดังกล่าว ข้อมูลที่ได้มีความสะอาด ถูกต้อง และอยู่ในรูปแบบที่พร้อมสำหรับการนำไปสร้างโมเดล Machine Learning หรือการทำ Exploratory Data Analysis ต่อไป

## Exploratory Data Analysis (EDA)

หลังจากการทำ Preprocessing ข้อมูลแล้ว จะนำข้อมูลที่ได้มาทำ Exploratory Data Analysis (EDA) ต่อ เพื่อทำความเข้าใจข้อมูลและเตรียมความพร้อมสำหรับการสร้างโมเดล Machine Learning ในขั้นตอนถัดไป

### 1. Import Libraries

มีการนำเข้าไลบรารีที่จำเป็นสำหรับการจัดการและวิเคราะห์ข้อมูล ได้แก่

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

เพื่อใช้ในการจัดการข้อมูล การแสดงผลในรูปแบบกราฟ และการวิเคราะห์ข้อมูลเชิงสถิติ

### 2. Load Dataset

ทำการโหลดชุดข้อมูลที่ผ่านการ Preprocessing แล้วจากไฟล์ CSV

```
df = pd.read_csv("final_cleaned_cpe_stress_data.csv")
```

### 3. ตรวจสอบและปรับปรุงข้อมูล

- ใช้ df.info() ตรวจสอบประเภทข้อมูล และ Missing Values
- เพิ่มคอลัมน์ SummaryST-5 ซึ่งรวมคะแนนจากแบบประเมินความเครียด ST-5 โดยใช้คำสั่ง

  ```
  df['SummaryST-5'] = df['1. มีปัญหาการนอน...'] + df['2. มีสมาธิน้อยลง...'] + ...
  ```

- ในขั้นตอนนี้ได้เกิดปัญหาขึ้นก็คือ ชื่อคอลัมน์มีทั้งภาษาไทยและภาษาอังกฤษ ทำให้ชื่อคอลัมน์มีความยาวมากเกินไป ส่งผลให้การวิเคราะห์ข้อมูลเป็นไปด้วยความยุ่งยาก จึงได้ทำการเปลี่ยนชื่อคอลัมน์ให้สั้นลงและเป็นภาษาอังกฤษ เพื่อให้ง่ายต่อการวิเคราะห์ข้อมูลในขั้นตอนถัดไป

### 4. วิเคราะห์ข้อมูล

- วิเคราะห์ข้อมูลเชิงสถิติด้วย df.describe()
- สร้างกราฟแจกแจงข้อมูล ด้วยการใช้คำสั่ง sns.histplot(df[col], kde=True)เพื่อให้สามารถมองเห็นข้อมูลได้ชัดเจนยิ่งขึ้น
- สร้างกราฟ ด้วยการใช้คำสั่ง sns.boxplot(data=df, x=col, y='SummaryST-5') เพื่อให้ง่ายต่อการสอบตรวจสอบการกระจายตัวของข้อมูล

- วิเคราะห์ความสัมพันธ์ระหว่างของข้อมูลด้วย heatmap ด้วยการใช้คำสั่ง sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')

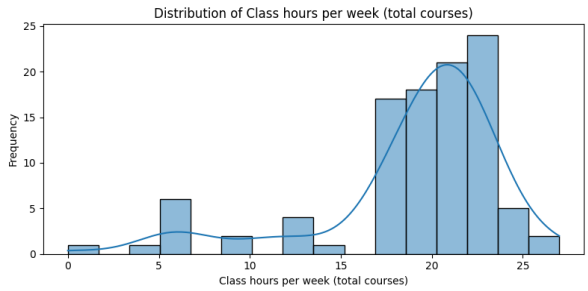
**5. สรุปข้อมูลที่ได้จากการ EDA**

- การวิเคราะห์ข้อมูลเชิงสถิติ ทำให้ทราบถึงค่าเฉลี่ย (Mean), ค่ามัธยฐาน (Median), ค่าสูงสุด/ต่ำสุด (Max/Min) การกระจายของข้อมูล ตัวอย่างเช่น

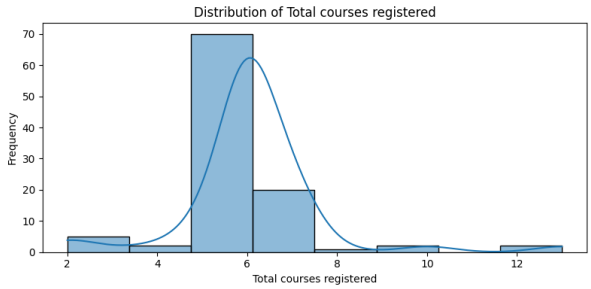
| columns | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Class hours per week (total courses) | 102 | 18.980 | 5.056 | 0 | 18 | 21 | 22 | 27 |
| Total courses registered | 102 | 6.127 | 1.5458 | 2 | 6 | 6 | 6 | 13 |
| Participation Hours per week | 102 | 5.852 | 10.014 | 0 | 0 | 2 | 5.75 | 41 |
| Extracurricular and self-study hours per week | 120 | 4.911 | 5.559 | 0 | 1 | 4 | 7 | 31 |
| SummaryST-5 | 102 | 7.509 | 3.013 | 0 | 5 | 8 | 10 | 15 |

*ตัวอย่างค่าทางสถิติที่ได้*

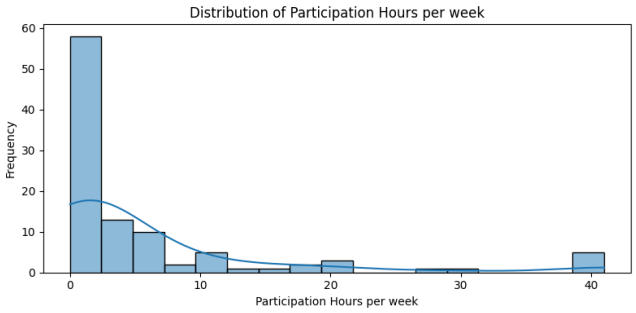
การแสดงผลในรูปแบบกราฟ ทำให้ทราบถึงข้อมูลโดยรวมของแต่ละคอลัมน์ ตัวอย่างเช่น



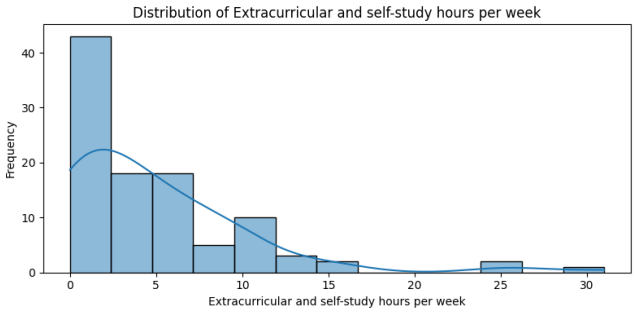

*กราฟแสดง Class hours per week (total courses)*




*กราฟแสดง Total courses registered*

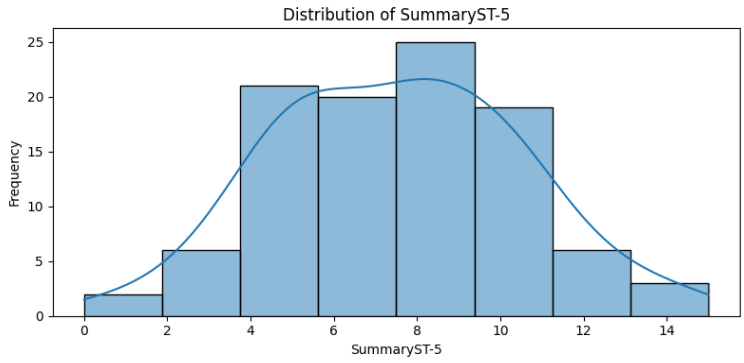



*กราฟแสดง Participation Hours per week*



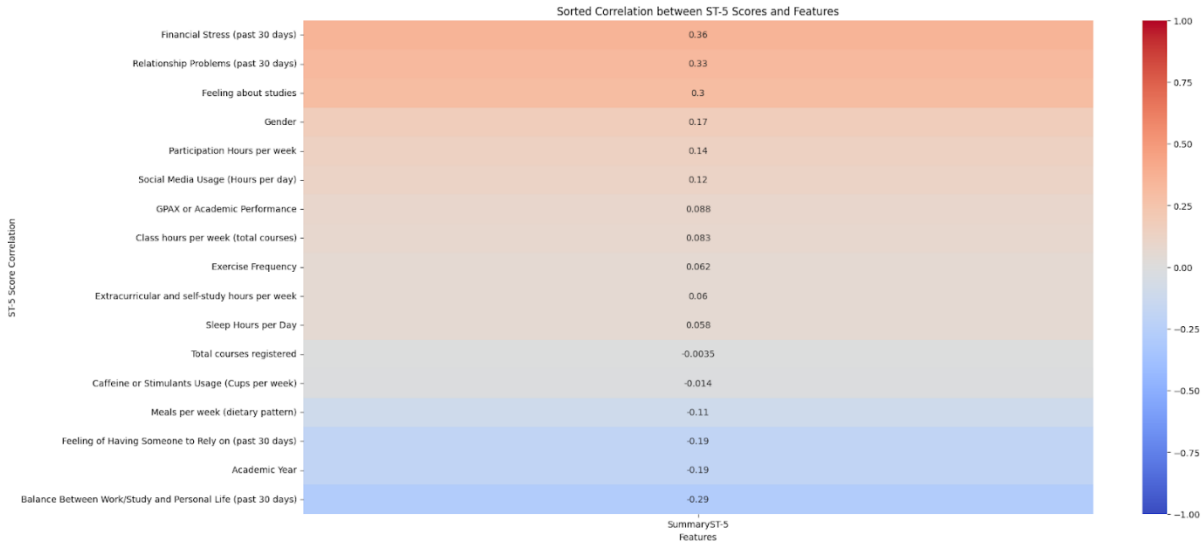

*กราฟแสดง Extracurricular and self-study hours per week*

- การสอบตรวจสอบการกระจายตัวของข้อมูล พบว่ามี outlier บางส่วน ซึ่งเป็นจำนวนที่น้อยมาก ซึ่งไม่มีผลต่อโมเดลที่เราใช้เทรนข้อมูล
- นอกจากนี้กราฟยังแสดงให้เห็นถึงแนวโน้มความเครียดของผู้ที่เข้ามาทำแบบสอบถาม ซึ่งแนวโน้มส่วนใหญ่จะเป็นความเครียดในระดับกลางไปถึงระดับสูง จึงต้องมีการ การสร้างโมเดล Machine Learning เพื่อศึกษาปัจจัยที่มีผลต่อความเครียด




*กราฟแสดง SummaryST-5*

- การวิเคราะห์ความสัมพันธ์ระหว่างของข้อมูล ทำให้ได้ค่า correlation มาดังนี้




*Heatmap ที่แสดงข้อมูลถึงความสัมพันธ์ของคอลัมน์ต่าง ๆ ต่อคอลัมน์ SummatyST-5*

นอกจากมีการวิเคราะห์อื่น ๆ ดังที่แสดงในภาคผนวก ข ได้แก่

- การกระจายตัวของชั้นปีการศึกษา
- การกระจายตัวของเกรดเฉลี่ยสะสม
- การกระจายตัวของเพศ
- ชั่วโมงการนอนหลับต่อวันเทียบกับคะแนน ST-5
- จำนวนวันที่ออกกำลังกายต่อสัปดาห์เทียบกับคะแนน ST-5
- ชั่วโมงการใช้โซเชียลมีเดียต่อวันเทียบกับคะแนน ST-5
- จำนวนมื้ออาหารที่รับประทานตามรูปแบบการควบคุมอาหารต่อสัปดาห์เทียบกับคะแนน ST-5

- คะแนนการประเมินตนเองของผู้ตอบแบบสอบถามช่วง 1 (คะแนนน้อยที่สุด) ถึง 10 (คะแนนมากที่สุด) เรื่องสมดุลระหว่างงาน/การเรียน กับชีวิตส่วนตัวช่วง 30 วันที่ผ่านมา เทียบกับคะแนน ST-5
- คะแนนการประเมินตนเองของผู้ตอบแบบสอบถามช่วง 1 (คะแนนน้อยที่สุด) ถึง 10 (คะแนนมากที่สุด) เรื่องการมีคนให้พึ่งพาได้ในช่วง 30 วันที่ เทียบกับคะแนน ST-5
- คะแนนการประเมินตนเองของผู้ตอบแบบสอบถามช่วง 1 (คะแนนน้อยที่สุด) ถึง 10 (คะแนนมากที่สุด) เรื่องความเครียดทางการเงินในช่วง 30 วันที่ผ่านมา เทียบกับคะแนน ST-5
- คะแนนการประเมินตนเองของผู้ตอบแบบสอบถามช่วง 1 (คะแนนน้อยที่สุด) ถึง 10 (คะแนนมากที่สุด) เรื่องการมีปัญหาความสัมพันธ์ในช่วง 30 วันที่ผ่านมา เทียบกับคะแนน ST-5

**6. ส่งต่อข้อมูล**

ส่งออกผลลัพธ์จากการวิเคราะห์เพื่อให้ง่ายต่อการนำไปใช้งานต่อในขั้นตอนถัดไป

- กราฟต่างๆ ถูกบันทึกเป็นไฟล์ .png
- ค่าทางสถิติจากคำสั่ง df.describe() ส่งออกเป็น .csv ด้วยคำสั่ง df.describe().to_csv('summary_statistics.csv')

**สรุปกระบวนการ EDA**

กระบวนการ EDA ที่ดำเนินการนี้ช่วยให้เข้าใจข้อมูลได้ดีขึ้น ช่วยให้สามารถคัดเลือกฟีเจอร์ที่เหมาะสม รวมทั้งเตรียมความพร้อมของข้อมูลเพื่อใช้ในการสร้างโมเดล Machine Learning อย่างมีประสิทธิภาพในขั้นตอนถัดไป

# Method

## Model Selection

1. โมเดลที่เลือกใช้

   Random Forest Classifier

2. เหตุผลที่เลือกใช้โมเดลนี้ เพราะ

   - Random Forest มีความสามารถในการจัดการกับข้อมูลที่มีหลายมิติ และลดความเสี่ยงจาก overfitting ได้ดี
   - สามารถจัดการกับปัญหาความไม่สมดุลของคลาส (class imbalance) ได้โดยใช้ class_weight=balanced
   - มีความยืดหยุ่นและประสิทธิภาพสูงสำหรับข้อมูลที่มีความซับซ้อน
   - มีการพิจารณา Logistic Regression และ Decision Tree ด้วย แต่ Random Forest มีความเหมาะสมมากกว่าจากการทดลองที่ผ่านมา

3. การปรับพารามิเตอร์ (Hyperparameter Tuning):

   - ใช้ Optuna เพื่อค้นหาค่าพารามิเตอร์ที่ดีที่สุด
   - พารามิเตอร์ที่ปรับ ได้แก่:

     o n_estimators

     o max_depth

     o min_samples_split

     o min_samples_leaf

     o max_features

     o bootstrap

     o class_weight

   - ใช้คะแนน F1-weighted ในการประเมินความแม่นยำระหว่าง tuning

## Training & Evaluation

1. วิธีการฝึกโมเดล:

   - ใช้ SMOTE เพื่อแก้ปัญหาคลาสไม่สมดุลก่อนแบ่งข้อมูล
   - แบ่งข้อมูลเป็น train/test (80/20) โดย stratified sampling

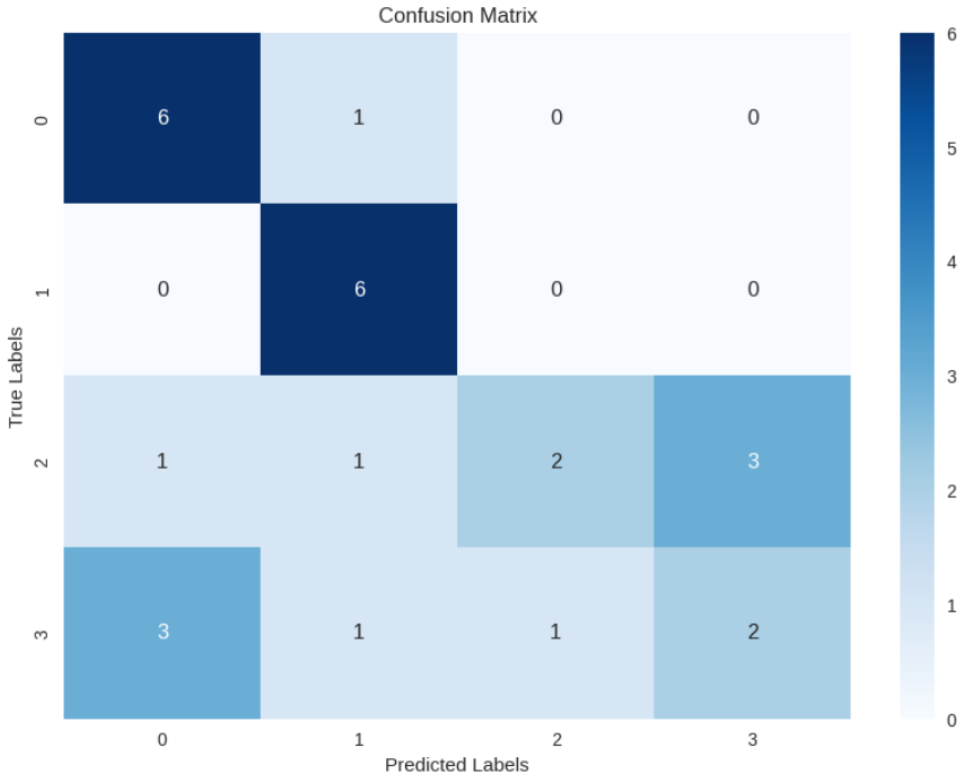
- ใช้ StratifiedKFold ร่วมกับ cross_val_score เพื่อประเมินแบบ cross-validation

2. เมตริกที่ใช้ในการประเมิน:Accuracy (ความแม่นยำ)
   - Precision (ทำนายว่าเป็นคลาสนั้น แล้วถูกจริงกี่เปอร์เซ็นต์)
   - Recall (จากทั้งหมดที่ควรเป็นคลาสนั้น โมเดลทำนายถูกกี่เปอร์เซ็นต์)
   - F1-score (ค่าเฉลี่ยถ่วงน้ำหนักระหว่าง precision กับ recall)
   - Confusion Matrix และ Classification Report
3. ผลลัพธ์ที่ได้จากโมเดล:
   - ได้ค่า F1-score สูงสุดจาก cross-validation โดยใช้ Optuna
   - บนชุดทดสอบ (Test Set):
     - Accuracy: (ค่าที่ print ออกมายังไม่ได้แสดงในผลลัพธ์ raw แต่สามารถเติมได้จากไฟล์จริง)
     - Precision, Recall, F1-score (แสดงผลด้วย classification_report)

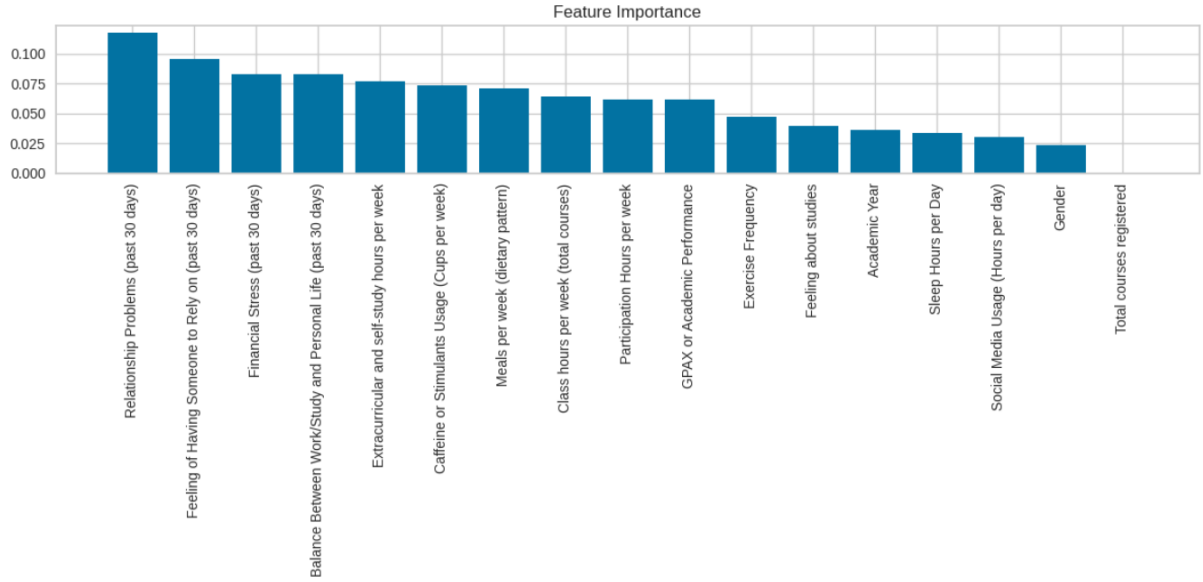
- แสดงผล Confusion Matrix และ Feature Importance

```
Test Set Evaluation:
Accuracy: 0.5925925925925926
Precision: 0.5802469135802469
Recall: 0.5925925925925926
F1 Score: 0.5509077705156137

Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.86      0.71         7
           1       0.67      1.00      0.80         6
           2       0.67      0.29      0.40         7
           3       0.40      0.29      0.33         7

    accuracy                           0.59        27
   macro avg       0.58      0.61      0.56        27
weighted avg       0.58      0.59      0.55        27
```

ภาพไฟล์ ผลลัพธ์การ tunning




ภาพผลลัพธ์ Confusion Matrix




ภาพผลลัพธ์ Feature Importance

**การวิเคราะห์ผล (Result Analysis)**

1. โมเดลทำงานดีในสถานการณ์ใด:

- เนื่องจากใช้ SMOTE, โมเดลมีความสามารถในการทำนายคลาสที่มีจำนวนน้อยได้ดีขึ้น
- การใช้ weighted metrics ช่วยให้การประเมินสอดคล้องกับคลาสไม่สมดุล

2. การตรวจสอบความลำเอียง (Bias) หรือความแปรปรวน (Variance):

- ใช้ cross-validation (StratifiedKFold) เพื่อลด bias จากการสุ่มข้อมูล
- Random Forest ช่วยลด variance จากการใช้โมเดล decision tree เดี่ยว

ข้อมูลเพิ่มเติม

- Optuna เป็น framework โอเพนซอร์สสำหรับการปรับไฮเปอร์พารามิเตอร์อัตโนมัติ ซึ่งออกแบบมาโดยเฉพาะสำหรับงานด้าน Machine Learning [1]
- SMOTE ย่อมาจาก Synthetic Minority Over-sampling Technique เป็นเทคนิคการโอเวอร์แซมปลิงที่ใช้เพื่อจัดการกับปัญหา ข้อมูลไม่สมดุล (imbalanced data) ในงานด้านการจำแนกประเภท classification ของ Machine Learning [2]

# Result & Discussion

## ผลการศึกษา (Results)

### ประสิทธิภาพของโมเดล

โมเดล Random Forest ที่พัฒนาขึ้นสำหรับการทำนายระดับความเครียดแสดงประสิทธิภาพในระดับปานกลาง:

- ความแม่นยำโดยรวม (Accuracy): 59.3%
- Weighted Precision: 61.1%
- Weighted Recall: 59.3%
- Weighted F1 Score: 58.7%

ประสิทธิภาพแยกตามระดับความเครียด:

- Low Stressed (0): Precision 75%, Recall 43%, F1-score 55%, Support 7
- Medium Stressed (1): Precision 56%, Recall 83%, F1-score 67%, Support 6
- High Stressed (2): Precision 57%, Recall 57%, F1-score 57%, Support 7
- Very High Stressed (3): Precision 57%, Recall 57%, F1-score 57%, Support 7

### ปัจจัยสำคัญที่มีอิทธิพลต่อระดับความเครียด

จากการวิเคราะห์คุณลักษณะด้วย Random Forest พบว่า 5 อันดับแรก ที่มีผลต่อการทำนายมากที่สุดคือ:

1. ความเครียดทางการเงิน (Financial Stress)
2. สมดุลระหว่างการเรียน/การทำงานและชีวิตส่วนตัว (Balance Between Work/Study and Personal Life)
3. จำนวนชั่วโมงเรียนต่อสัปดาห์ (Class hours per week)
4. ชั่วโมงกิจกรรมนอกหลักสูตรและการศึกษาด้วยตนเองต่อสัปดาห์ (Extracurricular and self-study hours)
5. มื้ออาหารต่อสัปดาห์/รูปแบบการรับประทานอาหาร (Meals per week)

ปัจจัยสำคัญระดับกลาง (อันดับ 6-10) ได้แก่:

- ปัญหาความสัมพันธ์ (Relationship Problems)
- การบริโภคคาเฟอีนหรือสารกระตุ้น (Caffeine or Stimulants Usage)
- ความรู้สึกว่ามีคนที่พึ่งพาได้ (Having Someone to Rely on)
- ชั่วโมงการมีส่วนร่วมในชั้นเรียนต่อสัปดาห์ (Participation Hours)
- ความรู้สึกเกี่ยวกับการเรียน (Feeling about studies)

ตัวแปรที่มีอิทธิพลน้อย (อันดับ 11-17) ได้แก่:

- การใช้สื่อสังคมออนไลน์ (Social Media Usage)
- ผลการเรียน (GPAX)
- จำนวนรายวิชาที่ลงทะเบียน (Total Courses)
- ความถี่ในการออกกำลังกาย (Exercise Frequency)
- ชั่วโมงการนอนต่อวัน (Sleep Hours)
- ชั้นปีการศึกษา (Academic Year)
- เพศ (Gender)

**ความสัมพันธ์กับคะแนนความเครียด**

จากการวิเคราะห์ความสัมพันธ์ (Correlation) กับคะแนน ST-5:

ความสัมพันธ์เชิงบวกที่แข็งแกร่งที่สุด (ปัจจัยที่สัมพันธ์กับความเครียดสูงขึ้น):

1. ความเครียดทางการเงิน (r = 0.36)
2. ปัญหาความสัมพันธ์ (r = 0.33)
3. ความรู้สึกเกี่ยวกับการเรียน (r = 0.30)

ความสัมพันธ์เชิงลบที่แข็งแกร่งที่สุด (ปัจจัยที่สัมพันธ์กับความเครียดต่ำลง):

1. สมดุลระหว่างการเรียน/การทำงานและชีวิตส่วนตัว (r = -0.29)
2. ชั้นปีการศึกษา (r = -0.19)
3. การมีคนที่พึ่งพาได้ (r = -0.19)
4. มื้ออาหารต่อสัปดาห์/รูปแบบการรับประทานอาหาร (r = -0.11)

## การอภิปรายผล (Discussion)

### การวิเคราะห์ Confusion Matrix

- กลุ่ม Medium Stressed (1) มีการทำนายที่แม่นยำที่สุด (5 จาก 6 กรณี)
- กลุ่ม Low Stressed (0) มีความท้าทายในการทำนายมากที่สุด โดยมีเพียง 3 จาก 7 กรณีที่ทำนายถูกต้อง
- มีความสับสนระหว่างกลุ่ม High Stressed (2) และ Very High Stressed (3) พอสมควร

### ความสัมพันธ์ระหว่างความสำคัญของคุณลักษณะและค่าสหสัมพันธ์

1. ความเครียดทางการเงินและสมดุลชีวิต
    - ความเครียดทางการเงินแสดงทั้งความสัมพันธ์สูงกับความเครียด (r = 0.36) และเป็นตัวทำนายที่สำคัญที่สุดในโมเดลการเรียนรู้ของเครื่อง
    - สมดุลชีวิตการทำงาน-การเรียนแสดงความสัมพันธ์เชิงลบที่แข็งแกร่ง (r = -0.29) และจัดอยู่ในอันดับที่สองของคุณลักษณะที่สำคัญที่สุด
    - ผลลัพธ์เหล่านี้บ่งชี้อย่างสม่ำเสมอว่าปัจจัยทางเศรษฐกิจสังคมและความสมดุลในการใช้ชีวิตมีความสำคัญมากกว่าปัจจัยทางวิชาการล้วนๆ ในการทำนายระดับความเครียด
2. ภาระงานทางวิชาการ vs. ผลการเรียน
    - น่าสนใจที่โมเดลระบุว่าจำนวนชั่วโมงเรียนและกิจกรรมนอกหลักสูตร/การศึกษาด้วยตนเองเป็นตัวทำนายที่สำคัญมาก แม้ว่าจะมีความสัมพันธ์ค่อนข้างน้อย
    - สิ่งนี้บ่งชี้ถึงความสัมพันธ์ที่ซับซ้อนและอาจไม่เป็นเชิงเส้นระหว่างการใช้เวลาและความเครียด
    - ในทางตรงกันข้าม ผลการเรียนจริง (GPAX) และชั้นปีการศึกษามีความสำคัญต่ำ แสดงว่าปริมาณงานมากกว่าคุณภาพของผลงานที่ขับเคลื่อนความเครียด
3. ปัจจัยด้านไลฟ์สไตล์และสุขภาพ
    - รูปแบบการรับประทานอาหาร (มื้ออาหารต่อสัปดาห์) จัดอันดับที่ห้าในด้านความสำคัญแม้จะมีความสัมพันธ์ปานกลาง (r = -0.11)
    - การบริโภคคาเฟอีน/สารกระตุ้นมีความสำคัญสูงกว่าที่คาดไว้เมื่อพิจารณาจากความสัมพันธ์ต่ำ

- การนอนหลับและการออกกำลังกายแสดงความสำคัญและความสัมพันธ์ต่ำกับความเครียดอย่างสม่ำเสมอ
- ผลลัพธ์เหล่านี้บ่งชี้ว่าปัจจัยด้านไลฟ์สไตล์บางประการอาจมีผลกระทบแบบเกณฑ์หรือมีปฏิสัมพันธ์กับตัวแปรอื่นๆ ในแบบที่ไม่ถูกจับโดยความสัมพันธ์อย่างง่าย

4. พลวัตการสนับสนุนทางสังคม
    - "การมีคนที่พึ่งพาได้" แสดงทั้งความสัมพันธ์ที่มีความหมาย (r = -0.19) และความสำคัญปานกลางของคุณลักษณะ
    - ปัญหาความสัมพันธ์แสดงความสัมพันธ์สูง (r = 0.33) และความสำคัญปานกลาง
    - สิ่งนี้ยืนยันความสำคัญของพลวัตทางสังคมในการจัดการความเครียดในกลุ่มนักศึกษา
5. ปัจจัยด้านประชากรศาสตร์และปัจจัยคงที่
    - เพศและชั้นปีการศึกษาจัดอันดับต่ำสุดในด้านความสำคัญ แสดงว่าปัจจัยที่ขับเคลื่อนความเครียดเป็นปัจจัยตามสถานการณ์มากกว่าประชากรศาสตร์
    - สิ่งนี้เน้นย้ำถึงประสิทธิผลที่อาจเกิดขึ้นของการแทรกแซงที่มีเป้าหมายโดยไม่คำนึงถึงลักษณะทางประชากรศาสตร์ของนักศึกษา

**นัยสำคัญสำหรับการแทรกแซงและมาตรการช่วยเหลือ**

1. การสนับสนุนด้านการเงินและการบริหารเวลา
    - ควรให้ความสำคัญกับโปรแกรมความช่วยเหลือทางการเงิน เนื่องจากความเครียดทางการเงินมีทั้งความสัมพันธ์สูงและเป็นตัวทำนายที่สำคัญที่สุด
    - การจัดอบรมเชิงปฏิบัติการด้านการบริหารเวลาที่เน้นสมดุลระหว่างการทำงานและชีวิตอาจมีประสิทธิภาพเป็นพิเศษ
    - การจัดการภาระงานทางวิชาการดูเหมือนจะมีความสำคัญมากกว่าการแทรกแซงผลการเรียน
2. บริการสนับสนุนแบบบูรณาการ
    - โมเดลชี้ให้เห็นว่าแนวทางแบบองค์รวมที่จัดการกับปัจจัยด้านการเงิน การบริหารเวลา และโภชนาการจะมีประสิทธิภาพมากที่สุด

- ความสำคัญสูงของปัจจัยหลายประการบ่งชี้ว่าการแทรกแซงที่เน้นเฉพาะด้านอาจมีผลกระทบจำกัด
- การระบุนักศึกษาที่มีปัจจัยเสี่ยงหลายประการตั้งแต่เนิ่นๆ อาจช่วยให้สามารถให้การสนับสนุนเชิงรุกได้

3. การทบทวนโปรแกรมวิชาการระดับภาควิชา
    - การประเมินชั่วโมงเรียนและการกระจายภาระงานอาจเป็นประโยชน์มากกว่าการมุ่งเน้นที่มาตรฐานผลการเรียน
    - ความสมดุลระหว่างเวลาเรียนในชั้นเรียนและข้อกำหนดการศึกษาด้วยตนเองควรได้รับความสนใจเป็นพิเศษ
    - อาจพิจารณาการปรับโครงสร้างตารางการศึกษาเพื่อสนับสนุนสมดุลระหว่างการทำงานและชีวิตที่ดีขึ้น
4. การนำโมเดลไปประยุกต์ใช้และข้อจำกัด
    - ประสิทธิภาพปานกลางของโมเดล (ความแม่นยำ 59.3%) ควรได้รับการพิจารณาเมื่อนำไปใช้ในระบบการทำนาย
    - ความแตกต่างระหว่างความสัมพันธ์และความสำคัญของคุณลักษณะสำหรับตัวแปรบางตัวบ่งชี้ถึงความสัมพันธ์ที่ซับซ้อน
    - การปรับปรุงโมเดลเพิ่มเติมอาจเน้นที่การสำรวจปฏิสัมพันธ์ระหว่างคุณลักษณะที่มีอันดับสูงสุด
    - พิจารณาแนวทางแบบผสมผสานที่ใช้ประโยชน์จากทั้งจุดแข็งของความสัมพันธ์และข้อมูลเชิงลึกแบบไม่เชิงเส้นจากโมเดล Random Forest

**ข้อจำกัดของการศึกษา**

**ขนาดตัวอย่างและกำลังทางสถิติ**

- ขนาดตัวอย่างที่ค่อนข้างเล็ก (102 คน) จำกัดกำลังทางสถิติของการวิเคราะห์และความสามารถในการนำผลการศึกษาไปใช้กับประชากรทั่วไป
- จากข้อมูลเพียง 102 รายการที่แบ่งออกเป็นสี่กลุ่มระดับความเครียด ทำให้แต่ละกลุ่มมีตัวแทนจำกัด
- ขนาดตัวอย่างที่เล็กอาจมีส่วนทำให้ความแม่นยำของโมเดลอยู่ในระดับปานกลาง (59.3%)

- การศึกษาในอนาคตควรมุ่งเน้นที่ตัวอย่างขนาดใหญ่ขึ้นเพื่อปรับปรุงความเสถียรและประสิทธิภาพของโมเดล

**ความไม่สมดุลของกลุ่มระดับความเครียด**

- มีความไม่สมดุลอย่างมีนัยสำคัญระหว่างระดับความเครียดในชุดข้อมูลของเรา:
    - ความเครียดปานกลาง (Medium Stressed): 33 คน (32.4%)
    - ความเครียดสูงมาก (Very High Stressed): 28 คน (27.5%)
    - ความเครียดสูง (High Stressed): 25 คน (24.5%)
    - ความเครียดต่ำ (Low Stressed): 16 คน (15.7%)
- การมีตัวแทนน้อยของนักศึกษาที่มีความเครียดต่ำ (เพียง 15.7% ของตัวอย่าง) อาจจำกัดความสามารถของโมเดลในการระบุปัจจัยที่เกี่ยวข้องกับความเครียดต่ำได้อย่างแม่นยำ
- แม้จะใช้ SMOTE สำหรับการ oversampling ในระหว่างการฝึกโมเดล แต่ความไม่สมดุลในชุดข้อมูลเดิมอาจมีอิทธิพลต่อการคำนวณความสำคัญของคุณลักษณะ
- ความไม่สมดุลนี้อาจสะท้อนการกระจายความเครียดที่แท้จริงในกลุ่มนักศึกษาวิศวกรรมคอมพิวเตอร์ แต่ก็อาจบ่งชี้ถึงอคติในการสุ่มตัวอย่างหรือการเลือกตัวเองในกลุ่มผู้ตอบแบบสอบถาม

**ความไม่สมดุลของการกระจายตัวของชั้นปีการศึกษา**

- จากการวิเคราะห์ของเรา การกระจายตัวของนักศึกษาตามชั้นปีมีความไม่สมดุลอย่างมีนัยสำคัญ:
    - ชั้นปีที่ 2: 63 คน (61.8%)
    - ชั้นปีที่ 1: 17 คน (16.7%)
    - ชั้นปีที่ 3: 15 คน (14.7%)
    - ชั้นปีที่ 4: 7 คน (6.9%)
- แม้ว่าชั้นปีการศึกษาจะมีความสำคัญต่ำในโมเดลของเรา แต่การกระจายตัวที่ไม่สม่ำเสมอของผู้ตอบแบบสอบถามระหว่างชั้นปีต่างๆ ยังคงเป็นข้อจำกัดด้านวิธีการ
- การมีตัวแทนมากเกินไปของชั้นปีที่ 2 (61.8%) อาจปิดบังปัจจัยความเครียดหรือกลไกการรับมือที่เฉพาะเจาะจงกับชั้นปี

- นักศึกษาชั้นปีที่ 1 อาจเผชิญกับปัจจัยความเครียดที่แตกต่างกัน (การปรับตัวกับชีวิตมหาวิทยาลัย) เมื่อเทียบกับนักศึกษาชั้นปีสุดท้าย (การหางาน ข้อกำหนดการสำเร็จการศึกษา)
- การวิเคราะห์เฉพาะชั้นปีถูกจำกัดโดยขนาดตัวอย่างที่ไม่เพียงพอภายในแต่ละกลุ่มชั้นปี

**ปัญหาคุณภาพข้อมูล**

- การมีค่า null เนื่องจากปัญหาในการออกแบบแบบสอบถามจำเป็นต้องมีการแทนที่ข้อมูลหรือลบแถว ซึ่งอาจทำให้เกิดอคติ
- ความผิดพลาดของมนุษย์ในการออกแบบแบบสำรวจเน้นย้ำถึงความสำคัญของการทดสอบแบบสอบถามอย่างละเอียดก่อนการนำไปใช้
- ความสามารถในการตรวจจับค่าผิดปกติหรือรูปแบบผิดปกติมีจำกัดเนื่องจากข้อจำกัดของขนาดตัวอย่าง

**ข้อจำกัดด้านวิธีการ**

- การออกแบบแบบตัดขวาง (cross-sectional) จับความเครียดเพียงจุดเดียวในช่วงเวลา ไม่สามารถแสดงความแปรปรวนตามเวลาของระดับความเครียดตลอดปีการศึกษา
- ข้อมูลที่รายงานด้วยตนเองในทุกตัวแปรอาจทำให้เกิดอคติในการรายงาน โดยเฉพาะสำหรับหัวข้อที่ละเอียดอ่อน เช่น ปัญหาความสัมพันธ์หรือผลการเรียน
- เครื่องมือประเมินความเครียด ST-5 แม้จะได้รับการตรวจสอบความถูกต้อง แต่ให้การวัดอย่างง่ายของปรากฏการณ์ทางจิตวิทยาที่ซับซ้อน
- การใช้ SMOTE oversampling มีความจำเป็นเพื่อแก้ไขความไม่สมดุลของคลาส แต่อาจแนะนำรูปแบบเทียมที่ไม่มีอยู่ในข้อมูลเดิม

**ข้อจำกัดเฉพาะของโมเดล**

- อัลกอริทึม Random Forest แม้จะมีประสิทธิภาพในการจับความสัมพันธ์แบบไม่เชิงเส้น แต่อาจ overfit กับชุดข้อมูลขนาดเล็ก
- ความแม่นยำปานกลางของโมเดล (59.3%) บ่งชี้ถึงความแปรปรวนที่ไม่ได้รับการอธิบายอย่างมากในระดับความเครียด
- ความสำคัญของคุณลักษณะใน Random Forests อาจได้รับอิทธิพลจากความสัมพันธ์ระหว่างตัวทำนาย ซึ่งอาจประเมินตัวแปรบางตัวสูงเกินไป

- การปรับ hyperparameter ด้วยข้อมูลจำกัดอาจไม่ได้ระบุพารามิเตอร์ที่เหมาะสมที่สุดอย่างแท้จริง

**ทิศทางการวิจัยในอนาคต**

1. การพัฒนาคุณลักษณะและศึกษาผลกระทบของการมีปฏิสัมพันธ์
    - สำรวจเงื่อนไขการมีปฏิสัมพันธ์ระหว่างตัวทำนายที่สำคัญ (เช่น ความเครียดทางการเงิน × สมดุลระหว่างการทำงานและชีวิต)
    - ตรวจสอบผลกระทบของเกณฑ์ที่อาจเกิดขึ้นในชั่วโมงเรียนและกิจกรรมนอกหลักสูตร
    - พัฒนาเกณฑ์วัดแบบผสมผสานที่อาจจับภาพความเป็นอยู่ที่ดีของนักศึกษาโดยรวมได้ดีขึ้น
2. พลวัตตามเวลาของความเครียด
    - การศึกษาระยะยาวเพื่อติดตามว่าปัจจัยความเครียดพัฒนาอย่างไรตลอดโปรแกรมการศึกษา
    - การตรวจสอบความแปรปรวนตามฤดูกาลในตัวทำนายความเครียด (ช่วงสอบ ช่วงต้นเทอมเทียบกับช่วงสิ้นเทอม)
3. การศึกษาประสิทธิผลของการแทรกแซง
    - การแทรกแซงที่มีเป้าหมายตามผลการค้นพบความสำคัญของคุณลักษณะระดับสูง
    - การศึกษาเปรียบเทียบของแนวทางการแทรกแซงปัจจัยเดียวเทียบกับหลายปัจจัย
    - มุ่งเน้นที่ปัจจัยที่มีผลกระทบสูงและอาจเปลี่ยนแปลงได้ เช่น ความรู้ทางการเงินและการบริหารเวลา
4. กลยุทธ์การปรับปรุงโมเดล
    - ทดลองกับวิธีการแบบ ensemble ที่รวมโมเดลตามความสัมพันธ์และโมเดลตามต้นไม้การตัดสินใจ
    - ทดสอบอัลกอริทึมทางเลือกที่อาจจับความสัมพันธ์ที่ซับซ้อนที่ระบุได้ดีขึ้น
    - ขยายชุดคุณลักษณะตามข้อมูลเชิงลึกจากโมเดลเริ่มต้นนี้

## Conclusion

จากการศึกษาเพื่อวิเคราะห์ปัจจัยที่มีผลต่อระดับความเครียดของนักศึกษาวิศวกรรมคอมพิวเตอร์ มจธ. พบว่าความเครียดของนักศึกษาไม่ได้เกิดจากปัจจัยใดปัจจัยหนึ่งเพียงอย่างเดียว แต่เป็นผลรวมของหลายมิติ ทั้งด้านเศรษฐกิจ สังคม การศึกษา และไลฟ์สไตล์ โดยเฉพาะความเครียดทางการเงินและความสมดุลระหว่างชีวิตส่วนตัวกับภาระทางการเรียนซึ่งเป็นปัจจัยที่มีอิทธิพลมากที่สุดต่อระดับความเครียด ผลการวิเคราะห์ข้อมูลเชิงสถิติร่วมกับโมเดล Random Forest ช่วยระบุลำดับความสำคัญของปัจจัยต่าง ๆ ได้อย่างชัดเจน และชี้ให้เห็นถึงความสำคัญของการสร้างสภาพแวดล้อมการเรียนรู้ที่สนับสนุนทั้งในด้านวิชาการและจิตใจ การค้นพบเหล่านี้สามารถนำไปประยุกต์ใช้ในการพัฒนานโยบายหรือมาตรการสนับสนุนนักศึกษาให้เหมาะสมยิ่งขึ้น ซึ่งอาจช่วยลดระดับความเครียดและส่งเสริมคุณภาพชีวิตของนักศึกษาได้อย่างยั่งยืนในอนาคต

## Reference

[1] Optuna Contributors, 2018, Optuna: A hyperparameter optimization framework, https://optuna.readthedocs.io/en/stable/, [8 พฤษภาคม 2568]

[2] Rukchanok Piyapanichayakul, ปัญหาข้อมูลไม่สมดุล (Imbalanced Data in Classification Model), https://ntcloudsolutions.ntplc.co.th/knowledge/imbalanced-data-classification/, [8 พฤษภาคม 2568]

**ภาคผนวก ก**

โครงสร้างและคำถามในแบบสอบถาม

ส่วนที่ 1: ข้อมูลประชากรศาสตร์ (Demographic Information)

1. รหัสนักศึกษา (Student ID)
2. เพศ (Gender)

ส่วนที่ 2: ปัจจัยทางการศึกษา (Academic Factors)

1. ผลการเรียนเฉลี่ยสะสม หรือผลการเรียนโดยรวม (GPAX or Academic Performance)
2. จำนวนชั่วโมงเรียนต่อสัปดาห์ของวิชาทั้งหมดที่ลงทะเบียนในภาคการศึกษานี้ (Class hours per week)
3. จำนวนรายวิชาทั้งหมดที่ลงทะเบียนในภาคการศึกษานี้ (Total number of courses)
4. ชั่วโมงที่ใช้ในการเข้าร่วมกิจกรรมนอกหลักสูตรต่อสัปดาห์ (Extracurricular activities - Hours per week)
5. จำนวนชั่วโมงต่อสัปดาห์ที่เรียนเสริมนอกหลักสูตรและเรียนเสริมด้วยตนเอง ไม่รวมช่วงสอบ (Self-study/extra study - Hours per week)
6. ในช่วงที่ผ่านมา คุณรู้สึกเกี่ยวกับการเรียนของตัวเองอย่างไรบ้าง (Feelings about your studies recently)

ส่วนที่ 3: พฤติกรรมการใช้ชีวิต (Lifestyle Habits)

1. ชั่วโมงการนอนต่อวัน (Sleep Hours per Day)
2. จำนวนมื้ออาหารที่รับประทานตามรูปแบบการควบคุมอาหารต่อสัปดาห์ (Meals per week following a dietary pattern)
3. ความยากลำบากในการจดจ่อกับความคิดของตัวเอง (Difficulty focusing thoughts)
4. ความถี่ของการอารมณ์แปรปรวน (Mood swings frequency)
5. ออกกำลังกายกี่วันต่อสัปดาห์ (Exercise frequency)
6. การบริโภคคาเฟอีนหรือสารกระตุ้น - จำนวนแก้วต่อสัปดาห์ (Caffeine/stimulants - Cups per week)

7. การใช้โซเชียลมีเดีย - ชั่วโมงต่อวัน (Social media usage - Hours per day)

ส่วนที่ 4: สุขภาพจิตและสังคม (Psychological & Social Well-being)

1. ระดับความเครียดในช่วง 30 วันที่ผ่านมา (Self-reported stress level)
2. อาการวิตกกังวลหรือซึมเศร้าในช่วง 30 วันที่ผ่านมา (Anxiety/depression symptoms)
3. ปัญหาในความสัมพันธ์ในช่วง 30 วันที่ผ่านมา (Relationship problems)
4. ความเครียดทางการเงินในช่วง 30 วันที่ผ่านมา (Financial stress)
5. คุณรู้สึกว่าคุณมีคนที่สามารถพึ่งพาได้ในช่วง 30 วันที่ผ่านมาเพียงใด (Support availability)
6. ในช่วง 30 วันที่ผ่านมา คุณสามารถจัดสมดุลระหว่างงาน/การเรียน กับชีวิตส่วนตัวได้ดีแค่ไหน (Work-life balance)

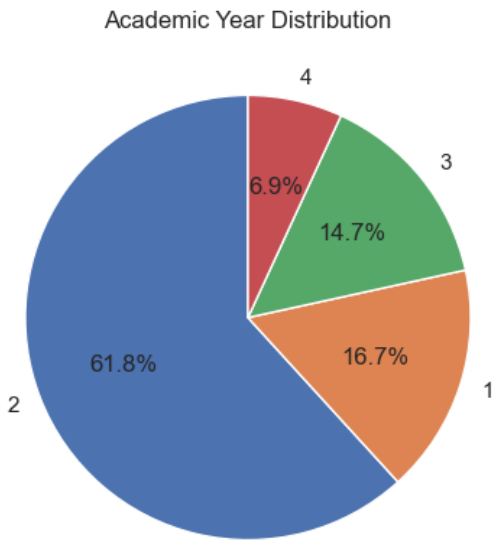
ส่วนที่ 5: ตัวแปรเป้าหมาย (Label)

1. มีปัญหาการนอน นอนไม่หลับหรือนอนมาก (Trouble sleeping)
2. มีสมาธิน้อยลง (Difficulty concentrating)
3. หงุดหงิด/กระวนกระวาย/ว้าวุ่นใจ (Irritability/restlessness)
4. รู้สึกเบื่อ เซ็ง (Boredom/lack of interest)
5. ไม่อยากพบปะผู้คน (Avoiding social interactions)

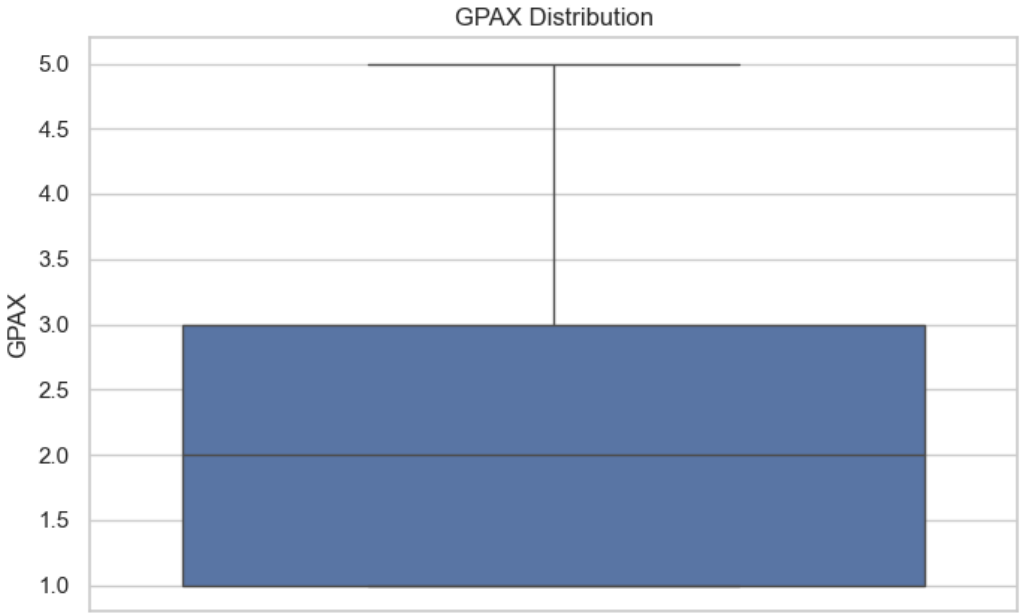
ลิ้งแบบสอบถาม

**ภาคผนวก ข**

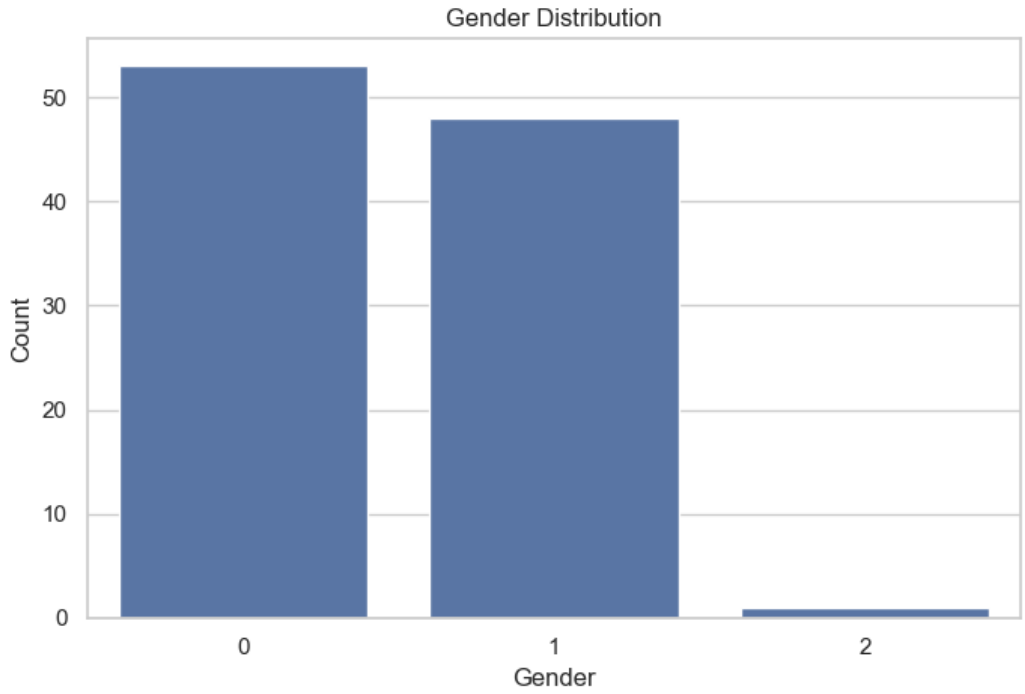
กราฟแสดงข้อมูลเพิ่มเติมในกระบวนการ EDA



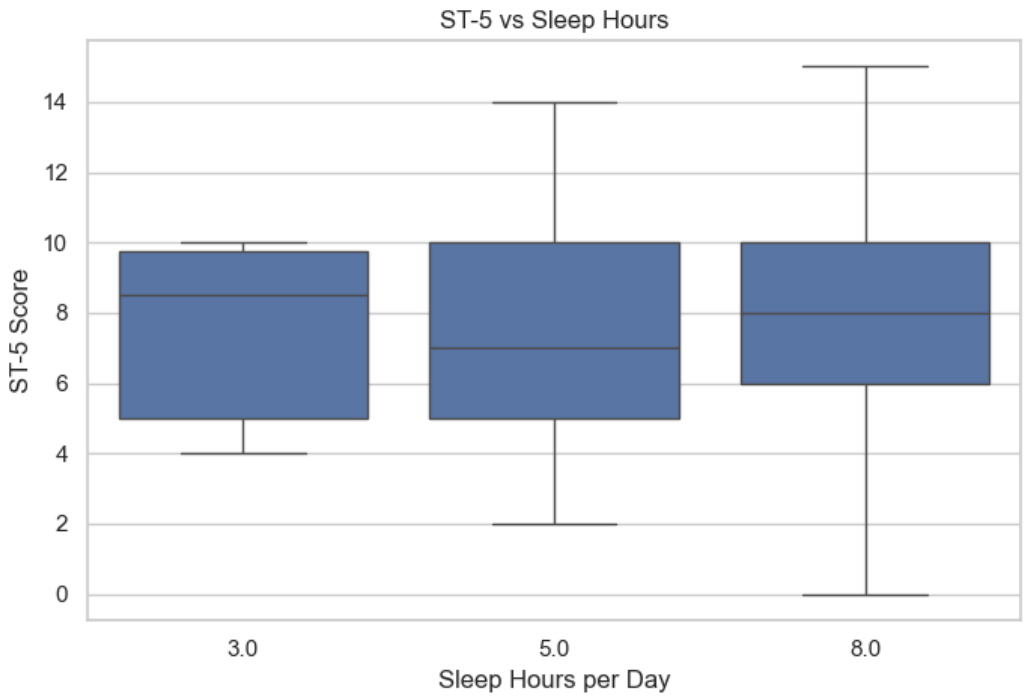

*การกระจายตัวของชั้นปีการศึกษา โดยเลข 1 - 2 หมายถึงนักศึกษาปี 1 - 4 ตามลำดับ*



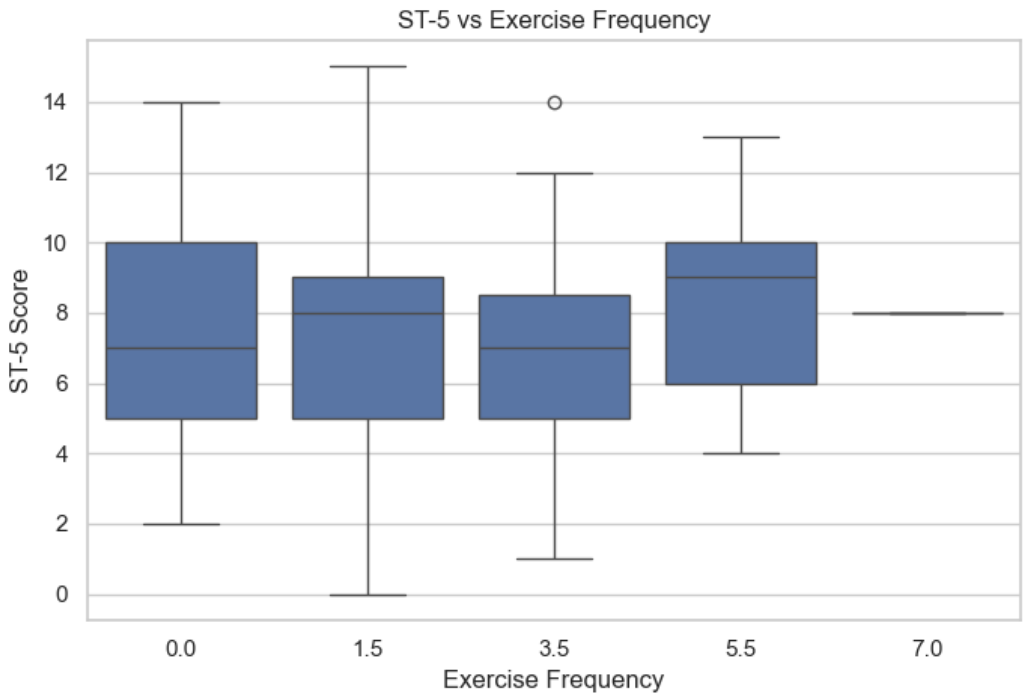

*การกระจายตัวของเกรดเฉลี่ยสะสม*



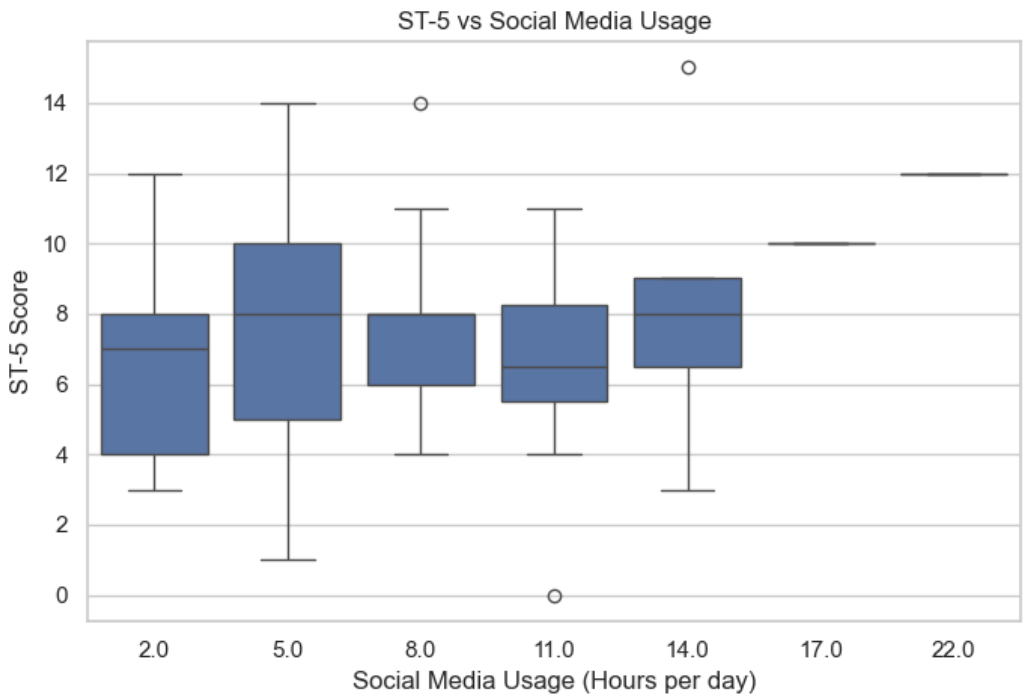

*การกระจายตัวของเพศ โดย 0 = ชาย, 1 = หญิง และ 2 = อื่น ๆ*



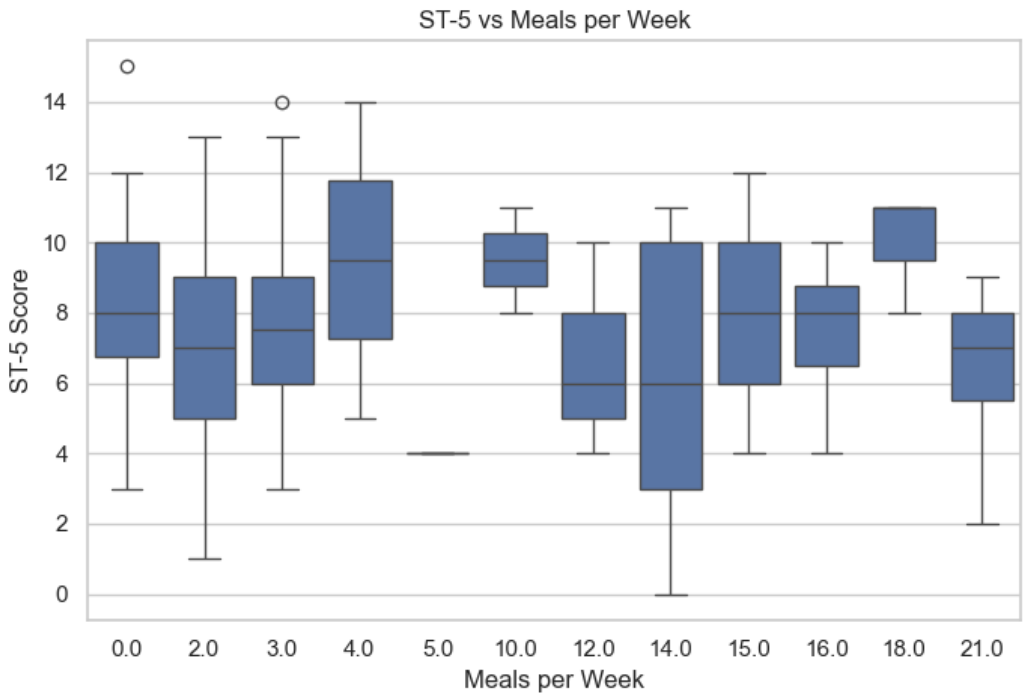

*ชั่วโมงการนอนหลับต่อวันเทียบกับคะแนน ST-5*



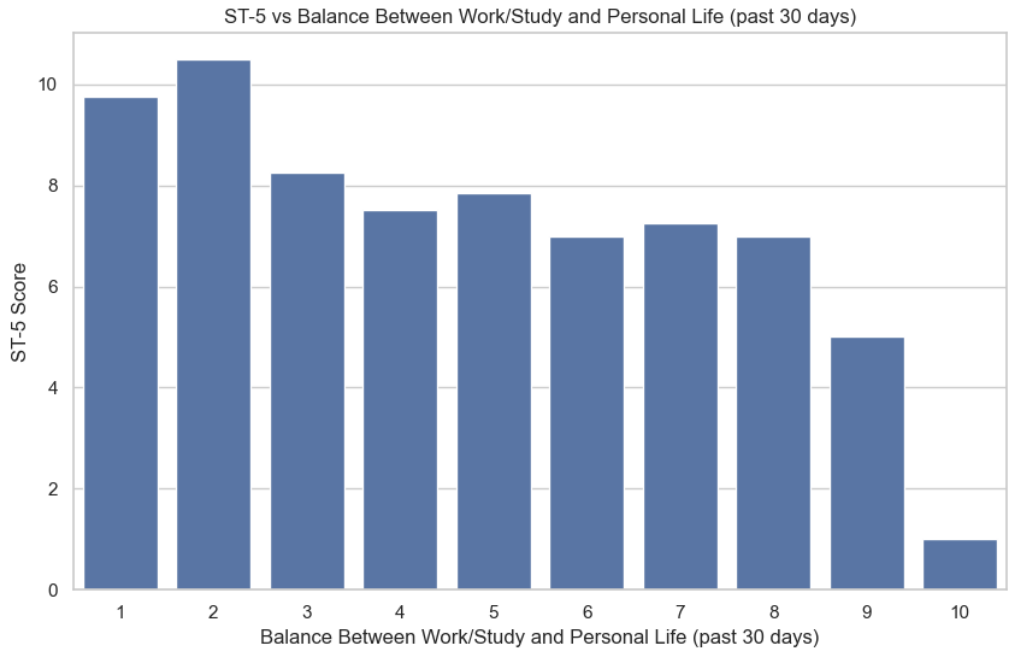

*จำนวนวันที่ออกกำลังกายต่อสัปดาห์เทียบกับคะแนน ST-5*



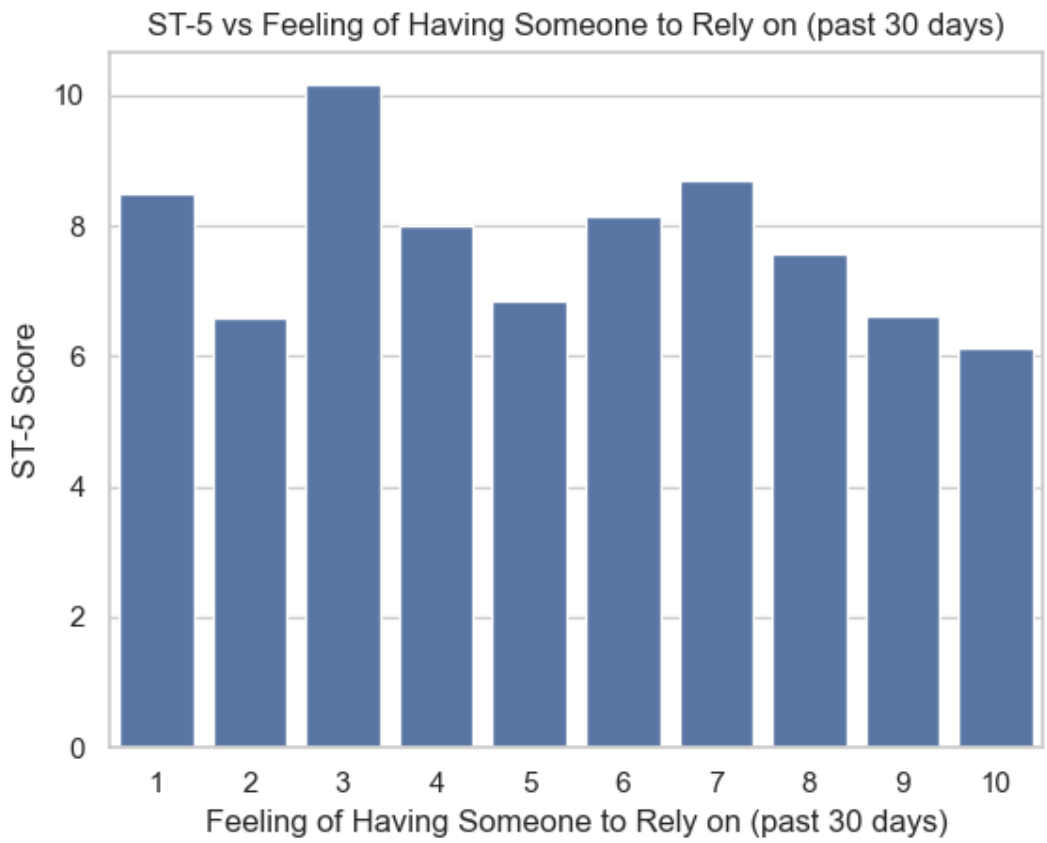

*ชั่วโมงการใช้โซเชียลมีเดียต่อวันเทียบกับคะแนน ST-5*




*จำนวนมื้ออาหารที่รับประทานตามรูปแบบการควบคุมอาหารต่อสัปดาห์เทียบกับคะแนน ST-5*

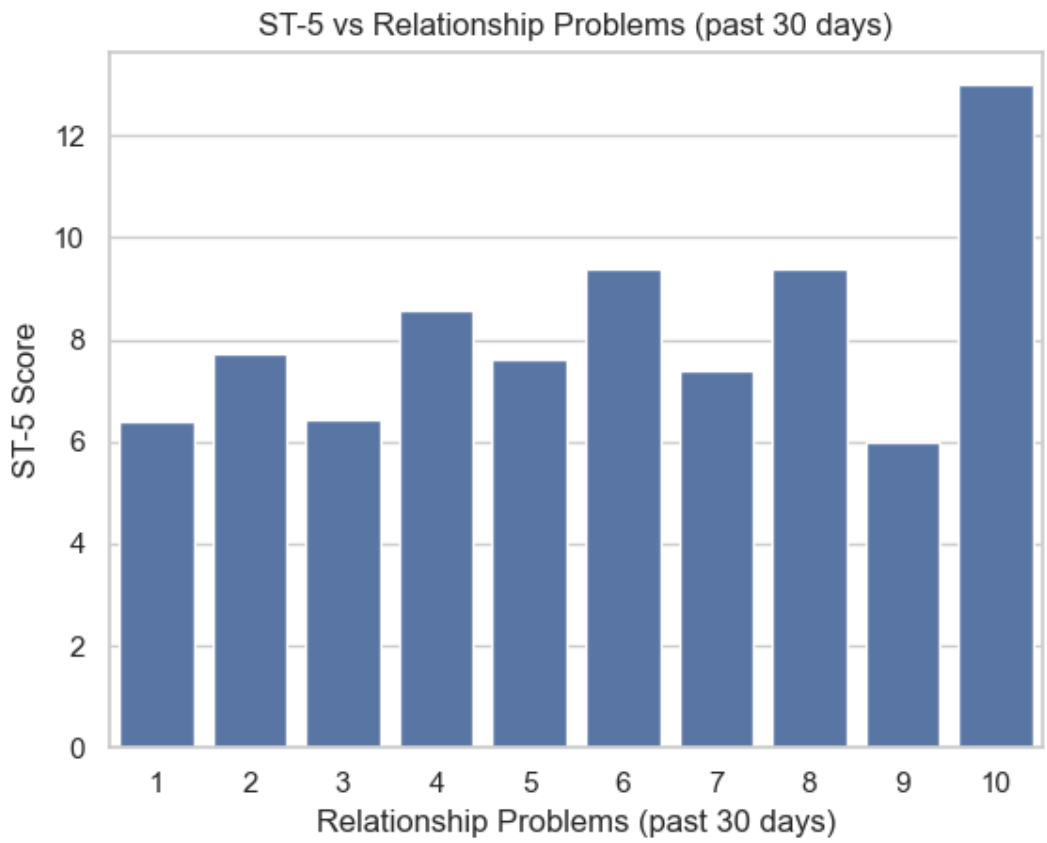



*คะแนนการประเมินตนเองของผู้ตอบแบบสอบถามช่วง 1 (คะแนนน้อยที่สุด) ถึง 10 (คะแนนมากที่สุด) เรื่อง สมดุลระหว่างงาน/การเรียน กับชีวิตส่วนตัวช่วง 30 วันที่ผ่านมา เทียบกับคะแนน ST-5*




*คะแนนการประเมินตนเองของผู้ตอบแบบสอบถามช่วง 1 (คะแนนน้อยที่สุด) ถึง 10 (คะแนนมากที่สุด) เรื่อง การมีคนให้พึ่งพาได้ในช่วง 30 วันที่ เทียบกับคะแนน ST-5*

*คะแนนการประเมินตนเองของผู้ตอบแบบสอบถามช่วง 1 (คะแนนน้อยที่สุด) ถึง 10 (คะแนนมากที่สุด) เรื่อง ความเครียดทางการเงินในช่วง 30 วันที่ผ่านมา เทียบกับคะแนน ST-5*




*คะแนนการประเมินตนเองของผู้ตอบแบบสอบถามช่วง 1 (คะแนนน้อยที่สุด) ถึง 10 (คะแนนมากที่สุด) เรื่องการมีปัญหาความสัมพันธ์ในช่วง 30 วันที่ผ่านมา เทียบกับคะแนน ST-5*