Final practice exam (20%)

การทดสอบนี้อนุญาตให้ใช้ Scikit-learn library, PyTorch library, หรือสร้างจากเริ่ม (Built from scratch) ได้

---

1. **เรื่อง: การวิเคราะห์การถดถอยเชิงเส้น (Regression analysis)**

   นักวิเคราะห์ต้องการสร้างโมเดลคำนวณเชิงเลขเพื่อหาความสัมพันธ์ของข้อมูลขาเข้ามีผลอย่างไรต่อผลกระทบต่อตลาดผู้บริโภคตาม **ตารางที่ 1** โดยมีข้อมูลฟีเจอร์ประกอบด้วย **(คะแนนรวมทั้งหมด 10%)**

   - ราคาเฉลี่ยของผลิตภัณฑ์ในกลุ่มเดียวกัน (Average Product Price) หน่วย: บาท/หน่วย
   - งบประมาณการโฆษณาต่อเดือน (Advertising Budget) หน่วย: ล้านบาท/เดือน
   - ความถี่ของการเปิดตัวสินค้าใหม่ (Product Launch Frequency) หน่วย: ครั้ง/ไตรมาส
   - คะแนนความพึงพอใจของลูกค้า (Customer Satisfaction Score) หน่วย: คะแนนเต็ม 100
   - ระดับผลกระทบต่อตลาดผู้บริโภค (Consumer Market Impact Level) หน่วย: คะแนนจาก 0 ถึง 10

   จงตอบคำถามต่อไปนี้

   1.1 จงสร้างโมเดลที่สามารถสร้างโมเดลเรียนรู้เชิงเลขและประเมินผลสัมประสิทธิ์การตัดสินใจระหว่างต้นแปรต้นต่อตัวแปรตามที่สูงที่สุด (Hint: Coefficient of Determination) **(คะแนน 8%)**

   1.2 สมมติข้อมูลทดสอบใหม่ เก็บผลทดสอบจากผลิตภัณฑ์หนึ่งใช้ราคาเฉลี่ยและงบประมาณในการโฆษณาน้อย แต่เน้นเพิ่มความถี่ในการเปิดตัวสินค้าใหม่มากทำให้ได้คะแนนความพึงพอใจมากตาม **(คะแนน 2%)**

   - ราคาเฉลี่ยของผลิตภัณฑ์ในกลุ่มเดียวกัน (Average Product Price) เท่ากับ 900 บาท/หน่วย
   - งบประมาณการโฆษณาต่อเดือน (Advertising Budget) เท่ากับ 0.5 ล้านบาท/เดือน
   - ความถี่ของการเปิดตัวสินค้าใหม่ (Product Launch Frequency) เท่ากับ 6 ครั้ง/ไตรมาส
   - คะแนนความพึงพอใจของลูกค้า (Customer Satisfaction Score) เท่ากับ 90 คะแนน

   ในขณะระดับผลกระทบต่อตลาดผู้บริโภค (Consumer Market Impact Level) ผลที่ได้ควรได้กี่คะแนนควรจะได้เท่าไหร่จาก 0 ถึง 10 (ผลกระทบจริงควรอยู่ในช่วง 7.8 ~ 8.0)

   <u>หมายเหตุ: สามารถเลือกใช้ตัวแปรต้นทั้งหมดหรือเลือกใช้ตัวแปรต้นตัวใดตัวหนึ่งก็ได้</u>

ตารางที่ 1 คลังข้อมูลข้อมูลผลกระทบต่อตลาดผู้บริโภค

ราคาเฉลี่ยของผลิตภัณฑ์ในกลุ่มเดียวกัน (Average Product Price) หน่วย: บาท/หน่วย	งบประมาณการโฆษณาต่อเดือน (Advertising Budget) หน่วย: ล้านบาท/เดือน	ความถี่ของการเปิดตัวสินค้าใหม่ (Product Launch Frequency) หน่วย: ครั้ง/ไตรมาส	คะแนนความพึงพอใจของลูกค้า (Customer Satisfaction Score) หน่วย: คะแนนเต็ม 100	ระดับผลกระทบต่อตลาดผู้บริโภค (Consumer Market Impact Level) หน่วย: คะแนนจาก 0 ถึง 10
ข้อมูลฝึกฝน (Train)				
1,500	2.5	1	85	7.2
1,200	1.8	2	78	6.5
1,800	3.2	1	90	8.1
900	1.0	0	60	4.3
1,300	2.0	2	75	6.8
2,000	4.0	3	88	9.0
1,600	2.8	2	82	7.5
1,000	1.2	1	70	5.2
1,400	2.3	2	77	6.9
1,700	3.5	3	91	8.7
ข้อมูลทดสอบ (Test)				
900	0.5	6	90	7.8 ~ 8.0

ตัวอย่างการประกาศข้อมูลใหม่และการแสดงผลพล็อต

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
# Data + user request
avg_product_price = [1500, 1200, 1800, 900, 1300, 2000, 1600, 1000, 1400, 1700]
ad_budget = [2.5, 1.8, 3.2, 1.0, 2.0, 4.0, 2.8, 1.2, 2.3, 3.5]
product_launch_freq = [1, 2, 1, 0, 2, 3, 2, 1, 2, 3]
customer_satisfaction_score = [85, 78, 90, 60, 75, 88, 82, 70, 77, 91]
consumer_market_impact_lvl = [7.2, 6.5, 8.1, 4.3, 6.8, 9.0, 7.5, 5.2, 6.9, 8.7]
# Create figure and axes for Size-Price plot
fig, (ax1, ax2, ax3, ax4) = plt.subplots(1, 4, figsize=(16, 6))

font_title = 9
ax1.scatter(ad_budget, consumer_market_impact_lvl, color='g', s=100,
edgecolors='black', alpha=0.7)
ax1.set_xlabel('ad_budget')
ax1.set_ylabel('consumer_market_impact_lvl')
ax1.set_title('ad_budget to consumer_market_impact_lvl', fontsize=font_title)
ax1.grid(True)
```

```
ax2.scatter(avg_product_price, consumer_market_impact_lvl, color='b', s=100,
edgecolors='black', alpha=0.7)
ax2.set_xlabel('avg_product_price')
ax2.set_ylabel('consumer_market_impact_lvl')
ax2.set_title('avg_product_price to consumer_market_impact_lvl',
fontsize=font_title)
ax2.grid(True)

ax3.scatter(product_launch_freq, consumer_market_impact_lvl, color='r', s=100,
edgecolors='black', alpha=0.7)
ax3.set_xlabel('product_launch_freq')
ax3.set_ylabel('consumer_market_impact_lvl')
ax3.set_title('product_launch_freq to consumer_market_impact_lvl',
fontsize=font_title)
ax3.grid(True)

ax4.scatter(customer_satisfaction_score, consumer_market_impact_lvl, color='y',
s=100, edgecolors='black', alpha=0.7)
ax4.set_xlabel('customer_satisfaction_score')
ax4.set_ylabel('consumer_market_impact_lvl')
ax4.set_title('customer_satisfaction_score to consumer_market_impact_lvl',
fontsize=font_title)
ax4.grid(True)

plt.tight_layout()
plt.show()

# *** Your code ***
###################################
# Create DataFrame …
# Data transformation …
# Train-test-sampling …
# Model initialization …
# Model fitting and evaluation metrics …
# Prediction …
```

2. **เรื่อง: งานการจำแนกประเภท (Classification task)**

บริษัทพลังงานแห่งหนึ่งในประเทศไทยต้องการสำรวจเครื่องปั่นไฟฟ้าที่พร้อมอยู่เพื่อเปิดใช้งานกังหันบางตัวที่ให้จ่ายพลังงานขาออกที่สูงต่อการใช้จ่ายโหลดต่อผู้บริโภคครัวเรือนตาม **ตารางที่ 2** โดยพิจารณาจากตัวแปรขาเข้าดังนี้ **(คะแนนรวมทั้งหมด 10%)**

- wind_speed: หมายถึงความเร็วลมที่ส่งผลโดยตรงต่อพลังงานที่ส่งออกจากกังหัน (5 ถึง 25 เมตร/วินาที)

- motor_tempreture: สามารถระบุสถานะของกังหันได้ หากกังหันมีอุณหภูมิสูง กังหันอาจต้องได้รับการซ่อมแซม (20 ถึง 80 องศาเซลเซียส)

- blade_angle: หากค่านี้เหมาะสมที่สุด จะช่วยให้ได้พลังงานที่ส่งออกเพิ่มขึ้น (0 ถึง 45 องศาเซลเซียส)

- vibration_level: ระดับสูงที่อาจบ่งชี้ถึงปัญหาทางกลไก (0 ถึง 10 ระดับ)

- humidity: ความชื้นสูงอาจส่งผลต่อประสิทธิภาพของกังหัน (20 ถึง 90%)

- air_pressure: ส่งผลต่อประสิทธิภาพของมุม (900 ถึง 1100 hPa)

ที่จำเป็นเพื่อทำนายพลังงานขาออก เกณฑ์การพิจารณาพลังงานที่ต้องการโดยนโยบายของบริษัทพลังงานนี้คือ

- 0-500 กิโลวัตต์ชั่วโมง (kWh) นับว่าเป็น "พลังงานขาออกน้อย"
- 500-600 กิโลวัตต์ชั่วโมง (kWh) นับว่าเป็น "พลังงานขาออกปานกลาง"
- 600 กิโลวัตต์ชั่วโมง (kWh) ขึ้นไป นับว่าเป็น "พลังงานขาออกสูง"

จงตอบคำถามต่อไปนี้

2.1 จงประมาณการโมเดลเรียนรู้ที่สามารถจำแนกประเภทพลังงานตามเกณฑ์ที่กำหนดและให้ความแม่นยำในการทำนายที่สูง **(คะแนน 7%)**

2.2 หากทีมงานทดสอบภาคสนามเก็บข้อมูลตัวอย่าง (สมมติคือข้อมูลแถวสุดท้ายที่ลบออกไป) เพื่อต้องการรู้ว่ากังหันลมต่อไปนี้ให้พลังงานเพียงพอที่ควรจะเปิดใช้งานหรือไม่ **(คะแนน 1.5%)**
   - แรงลม (wind_speed): 20 เมตร/วินาที
   - อุณหภูมิมอเตอร์ (motor_temperature): 80 องศาเซลเซียส
   - มุมของใบมีดใบพัดกังหันลม (blade_angle): 3 องศา
   - ระดับการสั่นสะเทือน (vibration_level): ระดับ 5
   - ความชื้น (humidity): 80%
   - ความกดอากาศ (air_pressure): 1000 hPa

2.3 จากข้อมูลเดียวกันกับข้อ 2.2 หากแรงลม (wind_speed) ลดลงจนเหลือ 0 เมตร/วินาที พลังงานที่ได้ควรเป็นอย่างไร **(คะแนน 1.5%)**

กำหนดให้ใช้ข้อมูลต่อไปนี้ทั้งหมดประกอบการทำข้อ 2

Dataset: “Kaggle - sajjadtahmasebi/wind-turbine-dataset”

Note: ให้ลบข้อมูลแถวสุดท้ายออกก่อน (แถวที่ 34999) แล้วทำการแบ่งข้อมูลด้วยการใช้ test size = ***20%*** และ random state สำหรับการสุ่มตัวอย่างที่ ***42***

ตารางที่ 2 ตัวอย่างข้อมูลฟีเจอร์สำหรับการวัดพลังงานขาออกจากกังหันลม

ใช้ฝึกและทดสอบสำหรับข้อ 2.1 เท่านั้น

	wind_speed	motor_temperature	blade_angle	vibration_level	humidity	air_pressure	energy_output	turbine_status
0	12.490802	56.563558	1.919750	8.618620	28.982391	963.680636	400.000000	needs to be repaired
1	24.014286	22.334208	37.282708	7.893506	74.230604	1057.022743	703.016603	off
2	19.639879	56.735620	11.218841	2.810425	63.106870	995.511060	465.597450	needs to be repaired
3	16.973170	25.380116	12.777181	4.197177	65.321011	1032.785998	418.536744	needs to be repaired
4	8.120373	62.620654	10.181038	5.076866	37.200880	934.721246	400.000000	needs to be repaired
...	...	...	...	...	...	...	...	...
34995	13.688071	55.663792	14.238366	7.877179	53.960014	1073.243533	400.000000	needs to be repaired
34996	14.271533	76.796947	8.328146	1.801420	79.767302	1093.872967	400.000000	needs to be repaired
34997	15.593338	75.255110	26.618267	5.302510	51.949326	1048.203685	400.000000	needs to be repaired
34998	24.351366	32.027862	6.324844	1.031443	57.138669	973.408237	586.548380	optimal
34999	23.093836	25.951137	14.852710	5.033187	85.185910	1020.801544	620.183282	needs to be repaired

35000 rows × 8 columns

ใช้ทดสอบในข้อ 2.2 และ 2.3 เท่านั้น

ตัวอย่างการกรองข้อมูลแถวสุดท้ายออกก่อน

```
# Wrangling
df.dropna(inplace=True)  # Drop rows with NaN
X = df.drop(['energy_output', 'turbine_status'], axis=1)[:-1]
y = df['energy_output'][:-1]

# *** Your code ***
###################################
# …
# …
# …
# …
```