ชื่อ - นามสกุล.....รหัสนักศึกษา.....

ข้อที่ 1: กระบวนการจัดการข้อมูลและการตีความ (Data Process & Interpretation)

(รวม 15 คะแนน | ระดับความยาก: ปานกลาง)

สถานการณ์: สมมติว่าท่านเป็นวิศวกรคอมพิวเตอร์ในทีมวิเคราะห์ข้อมูล และได้รับชุดข้อมูลดิบ (Raw Data) เกี่ยวกับประสิทธิภาพของนักกีฬา ซึ่งมีข้อมูล อายุ (Age) และ เงินเดือน (Salary) แต่ข้อมูลชุดนี้มีปัญหาดังนี้:

- มีค่าที่ขาดหาย (Missing Values) ในคอลัมน์ อายุ
- คอลัมน์ เงินเดือน มีหน่วยเป็นดอลลาร์สหรัฐ ซึ่งมีค่าสูงมากเมื่อเทียบกับอายุ (เช่น อายุ 25 ปี, เงินเดือน 5,000,000 ดอลลาร์)

เพื่อนร่วมงานของท่านเสนอว่า: "เพื่อความรวดเร็ว เรานำข้อมูล อายุ และ เงินเดือน ไปสร้าง Scatter Plot ด้วย Seaborn ทันทีเลย จะได้เห็นความสัมพันธ์เบื้องต้น"

คำถาม: ในฐานะวิศวกรคอมพิวเตอร์ จงประเมินข้อเสนอของเพื่อนร่วมงาน โดยพิจารณาทั้งข้อดีและข้อเสียที่อาจเกิดขึ้นจากการดำเนินการตามแนวทางดังกล่าวทันที จากนั้นให้สรุปแนวทางการดำเนินงานที่ท่านคิดว่าเหมาะสมที่สุด พร้อมอธิบายเหตุผลประกอบอย่างละเอียด โดยอ้างอิงถึงหลักการที่ได้เรียนมาในสัปดาห์ที่ 3 (Pandas) และสัปดาห์ที่ 4 (Data Visualization)

- **(5 คะแนน)** วิเคราะห์ข้อดีและข้อเสียของแนวทางที่เพื่อนร่วมงานเสนอ
- **(10 คะแนน)** เสนอแนวทางการดำเนินงานที่ท่านคิดว่าเหมาะสมที่สุด และอธิบายเหตุผลโดยละเอียดว่าทำไมแนวทางนั้นจึงดีกว่า โดยเชื่อมโยงกับลักษณะของข้อมูลที่ให้มาและผลกระทบต่อการวิเคราะห์

ชื่อ - นามสกุล.....รหัสนักศึกษา.....

ข้อที่ 2: การประเมินผลแบบจำลองในบริบทที่สำคัญ (Model Evaluation in a Critical Context)

(รวม 20 คะแนน | ระดับความยาก: สูง)

สถานการณ์: โรงพยาบาลแห่งหนึ่งได้พัฒนาแบบจำลอง K-Nearest Neighbors (k-NN) เพื่อช่วยแพทย์คัดกรองภาพถ่ายเนื้อเยื่อว่าเป็น "เนื้อร้าย (Malignant)" หรือ "เนื้อดี (Benign)" หลังจากทดสอบแบบจำลองกับข้อมูลชุดทดสอบ (Test Set) พบว่ามีค่า Accuracy (ความแม่นยำโดยรวม) สูงถึง 97% ผู้บริหารโรงพยาบาลพอใจกับตัวเลขนี้มาก แต่ทีมแพทย์ผู้เชี่ยวชาญยังคงกังวล

คำถาม: จงใช้ความรู้เรื่อง Confusion Matrix, Precision, และ Recall ที่เรียนในสัปดาห์ที่ 5 มาวิเคราะห์สถานการณ์นี้และตอบคำถามต่อไปนี้:

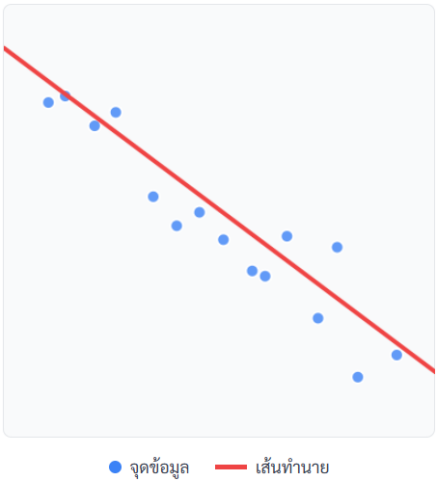
1. **(6 คะแนน)** เหตุใดค่า Accuracy ที่สูงถึง 97% อาจยังไม่เพียงพอที่จะทำให้แพทย์เชื่อมั่นในแบบจำลองนี้ได้?
2. **(8 คะแนน)** ในบริบททางการแพทย์นี้ ระหว่าง False Positive (FP) และ False Negative (FN) การทายผลผิดพลาดแบบใดที่ส่งผลกระทบร้ายแรงกว่ากัน? จงอธิบายผลกระทบที่เกิดขึ้นจริงของข้อผิดพลาดแต่ละแบบ
3. **(6 คะแนน)** เพื่อให้แบบจำลองนี้มีประโยชน์สูงสุดทางการแพทย์ ทีมแพทย์ควรให้ความสำคัญกับค่า Precision หรือ Recall ของคลาส "เนื้อร้าย (Malignant)" มากกว่ากัน? เพราะเหตุใด?

ชื่อ - นามสกุล.....รหัสนักศึกษา.....

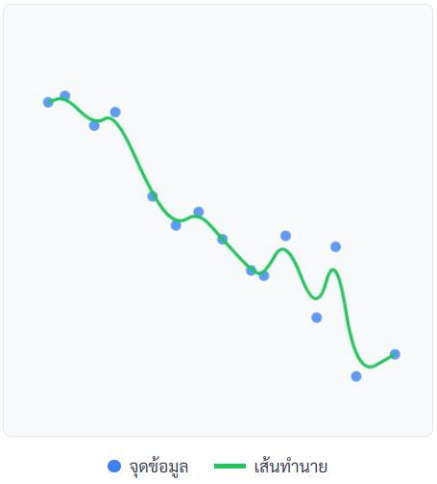
ข้อที่ 3: การวินิจฉัยประสิทธิภาพของแบบจำลอง (Overfitting & Underfitting)

(รวม 15 คะแนน | ระดับความยาก: ปานกลาง)

สถานการณ์: ท่านกำลังสร้างแบบจำลองเพื่อทำนายราคาบ้านจากขนาดพื้นที่ และได้ผลลัพธ์ของแบบจำลอง 2 แบบดังภาพ:



แบบจำลอง A: สร้างเส้นทำนายเป็นเส้นตรง (Linear Regression)



แบบจำลอง B: สร้างเส้นทำนายเป็นเส้นโค้งที่ซับซ้อนและพยายามวิ่งผ่านทุกจุดข้อมูล (High-degree Polynomial Regression)

ชื่อ - นามสกุล.....รหัสนักศึกษา.....

คำถาม:

1. **(5 คะแนน)** แบบจำลองใดมีแนวโน้มที่จะเกิดปัญหา Underfitting และแบบจำลองใดมีแนวโน้มที่จะเกิดปัญหา Overfitting? จงให้เหตุผล
2. **(5 คะแนน)** อธิบายว่า "คะแนนตอนสอน (Training Score)" และ "คะแนนตอนทดสอบ (Test Score)" ของแบบจำลอง B จะมีลักษณะเป็นอย่างไรเมื่อเทียบกัน?
3. **(5 คะแนน)** จากสถานการณ์เดียวกัน ปัญหา Underfitting ในแบบจำลอง A และ Overfitting ในแบบจำลอง B สะท้อนให้เห็นถึงปัญหาเรื่องความเอนเอียง (Bias) และความแปรปรวน (Variance) อย่างไร? จงอธิบายว่าแบบจำลองใดมี High Bias และแบบจำลองใดมี High Variance

ชื่อ - นามสกุล.....รหัสนักศึกษา.....

ข้อที่ 4: การคำวณและวิเคราะห์ k-Nearest Neighbors (k-NN)

(รวม 20 คะแนน | ระดับความยาก: สูง)

สถานการณ์: กำหนดให้มีชุดข้อมูล 2 มิติ (Feature 1, Feature 2) และ 2 คลาส (A, B) ดังตาราง และมี จุดข้อมูลใหม่ (P_new) ที่ตำแหน่ง (X=5, Y=3)

จุดข้อมูล	Feature 1 (X)	Feature 2 (Y)	คลาส
P1	2	2	A
P2	3	4	A
P3	5	5	A
P4	6	2	B
P5	8	3	B
P6	7	4	B

คำสั่ง:

1. **(8 คะแนน)** จงคำนวณระยะห่างแบบยูคลิด (Euclidean Distance) จากจุด P_new ไปยังจุดข้อมูลอื่นๆ ทั้ง 6 จุด (แสดงวิธีทำหรือตารางผลลัพธ์)

2. **(6 คะแนน)** จากผลการคำนวณในข้อ **(1.)** จงทำนายว่า P_new ควรจะอยู่ในคลาสใด เมื่อกำหนดให้ **k = 3**? พร้อมอธิบายขั้นตอนการลงคะแนน (Voting)

3. **(6 คะแนน)** หากเปลี่ยนค่า k เป็น 5 ผลการทำนายจะยังเหมือนเดิมหรือไม่? และการเพิ่มค่า k โดยทั่วไปส่งผลต่อความซับซ้อนของขอบเขตการตัดสินใจ (Decision Boundary) อย่างไร?

ชื่อ - นามสกุล.....รหัสนักศึกษา.....

ข้อที่ 5: การคำนวณและวิเคราะห์เมตริกสำหรับประเมินผล (Evaluation Metrics)

(รวม 15 คะแนน | ระดับความยาก: ปานกลาง)

สถานการณ์: หลังจากสร้างแบบจำลองสำหรับคัดกรองอีเมลสแปม ท่านได้ผลลัพธ์เป็น Confusion Matrix ดังนี้:

	ทำนายว่า: Not Spam	**ทำนายว่า: Spam**
ค่าจริง: Not Spam	150 (TN)	10 (FP)
ค่าจริง: Spam	20 (FN)	70 (TP)

คำสั่ง:

จากตาราง Confusion Matrix ข้างต้น จงคำนวณค่าต่อไปนี้ (แสดงสูตรและวิธีทำ):

1. **(3 คะแนน)** Accuracy

2. **(3 คะแนน)** Precision ของคลาส "Spam"

3. **(3 คะแนน)** Recall ของคลาส "Spam"

4. **(6 คะแนน)** ในบริบทของการคัดกรองอีเมลสแปมนี้ ระหว่าง Precision และ Recall ท่านคิดว่าเมตริกใดมีความสำคัญมากกว่ากัน? การที่ค่าใดค่าหนึ่งต่ำจะส่งผลเสียต่อประสบการณ์ของผู้ใช้อย่างไร?

ชื่อ - นามสกุล.....รหัสนักศึกษา.....

ข้อที่ 6: การคำนวณและวิเคราะห์ Linear Regression

(รวม 15 คะแนน | ระดับความยาก: ปานกลาง)

สถานการณ์: กำหนดให้มีชุดข้อมูลขนาดพื้นที่และราคาคอนโดดังนี้:

ขนาด (ตร.ม.), X	ราคา (ล้านบาท), Y
30	2.0
50	3.0
80	5.0

คำสั่ง:

1. **(7 คะแนน)** จากข้อมูลที่กำหนดให้ จงคำนวณหาค่าความชัน (m) และจุดตัดแกน (c) ของสมการเส้นตรง y=mx+c (แสดงวิธีทำ)

2. **(4 คะแนน)** จงเขียนสมการเส้นตรงที่ได้จากข้อ ก. และใช้สมการดังกล่าวทำนายราคาคอนโดที่มีขนาด 60 ตารางเมตร

3. **(4 คะแนน)** จากสถานการณ์นี้ แบบจำลอง Linear Regression ที่ท่านสร้างขึ้นมีแนวโน้มที่จะมี High Bias หรือ High Variance มากกว่ากัน? และเป็นเพราะเหตุใด?