**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

**ข้อที่ 1: กระบวนการจัดการข้อมูลและการตีความ (Data Process & Interpretation)**

**(1) วิเคราะห์ข้อดีและข้อเสียของแนวทางที่เพื่อนร่วมงานเสนอ**

**ข้อเสนอของเพื่อนร่วมงานที่ให้สร้าง Scatter Plot ทันทีมีทั้งข้อดีและข้อเสียดังนี้**

* **ข้อดี:**
  + **ความรวดเร็ว: เป็นวิธีที่ง่ายและรวดเร็วในการเห็นภาพรวมความสัมพันธ์เบื้องต้นระหว่าง "อายุ" และ "เงินเดือน"**
* **ข้อเสีย:**
  + **การจัดการค่าที่ขาดหาย (Missing Values): การพล็อตข้อมูลที่มีค่าขาดหายในคอลัมน์ "อายุ" อาจทำให้เกิดข้อผิดพลาด (Error) หรือทำให้ไลบรารีอย่าง Seaborn อาจตัดแถวข้อมูลนั้นทิ้งไปโดยอัตโนมัติ ซึ่งส่งผลให้การตีความความสัมพันธ์ผิดเพี้ยนไปจากความเป็นจริง**
  + **ปัญหาเรื่องสเกล (Scale) ของข้อมูล: คอลัมน์ "เงินเดือน" มีค่าสูงมากเมื่อเทียบกับ "อายุ" การนำไปพล็อตโดยตรงจะทำให้แกน Y (เงินเดือน) มีช่วงกว้างมาก ในขณะที่แกน X (อายุ) มีช่วงแคบ การกระจายตัวของจุดข้อมูลอาจจะถูกบีบอัดตามแนวแกน X ทำให้มองเห็นแนวโน้มหรือรูปแบบความสัมพันธ์ได้ไม่ชัดเจน**
  + **การตีความที่ผิดพลาด: จากปัญหาทั้งสองข้อข้างต้น Scatter Plot ที่ได้อาจให้ภาพที่ไม่ถูกต้องเกี่ยวกับความสัมพันธ์ที่แท้จริงระหว่างตัวแปร นำไปสู่การตัดสินใจที่ผิดพลาดในขั้นตอนต่อไป**

**(2) เสนอแนวทางการดำเนินงานที่เหมาะสมที่สุด**

**แนวทางการดำเนินงานที่เหมาะสมที่สุดควรเป็นไปตามลำดับขั้นตอนการเตรียมข้อมูล (Data Preprocessing) ก่อนนำไปสร้างภาพ (Data Visualization) ดังนี้**

1. **การจัดการค่าที่ขาดหาย (Handling Missing Values):**
   * **ขั้นตอน: ใช้ไลบรารี Pandas เพื่อตรวจสอบหาค่าที่ขาดหายในคอลัมน์ "อายุ" จากนั้นตัดสินใจเลือกวิธีจัดการที่เหมาะสม เช่น การเติมค่าที่ขาดหายด้วยค่าเฉลี่ย (Mean) หรือค่ามัธยฐาน (Median) ของอายุ ซึ่งมักเป็นตัวเลือกที่ดีสำหรับข้อมูลเชิงปริมาณ**
   * **เหตุผล: การเติมค่าที่ขาดหายช่วยให้เราไม่สูญเสียข้อมูลในแถวนั้นๆ ไป ทำให้สามารถนำข้อมูลไปใช้ในการวิเคราะห์ได้อย่างครบถ้วนและแม่นยำยิ่งขึ้น**
2. **การปรับสเกลข้อมูล (Data Scaling/Normalization):**
   * **ขั้นตอน: เนื่องจาก "เงินเดือน" มีค่าสูงกว่า "อายุ" มาก ควรทำการปรับสเกลข้อมูลเพื่อให้ตัวแปรทั้งสองมีช่วงค่าที่ใกล้เคียงกัน อาจใช้วิธี Min-Max Normalization (ปรับค่าให้อยู่ในช่วง 0 ถึง 1) หรือ Standardization (ปรับให้มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1)**
   * **เหตุผล: การปรับสเกลช่วยให้การสร้างภาพและการวิเคราะห์ด้วยแบบจำลองบางประเภท (เช่น k-NN) มีประสิทธิภาพมากขึ้น เพราะป้องกันไม่ให้ตัวแปรที่มีค่าสูงกว่ามีอิทธิพลต่อการวิเคราะห์มากเกินไป และทำให้ Scatter Plot แสดงการกระจายตัวของข้อมูลได้ชัดเจนยิ่งขึ้น**
3. **การสร้างภาพข้อมูล (Data Visualization):**
   * **ขั้นตอน: หลังจากเตรียมข้อมูลเสร็จสิ้นแล้ว จึงนำข้อมูลที่ผ่านการจัดการค่าขาดหายและปรับสเกลแล้วไปสร้าง Scatter Plot ด้วย Seaborn**
   * **เหตุผล: การทำเช่นนี้จะทำให้ได้กราฟที่สะท้อนความสัมพันธ์ที่แท้จริงระหว่างตัวแปรได้ดีที่สุด ปราศจากปัญหาที่เกิดจากข้อมูลดิบ ทำให้การตีความผลลัพธ์น่าเชื่อถือและนำไปสู่การสร้างแบบจำลองที่แม่นยำต่อไป**

**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?

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

**(1) เหตุใดค่า Accuracy ที่สูงถึง 97% อาจยังไม่เพียงพอ**

**ค่า Accuracy 97% อาจยังไม่เพียงพอที่จะทำให้แพทย์เชื่อมั่นในบริบทของการคัดกรองเนื้อร้าย เพราะค่า Accuracy บอกเพียงภาพรวมว่าแบบจำลองทายถูกกี่เปอร์เซ็นต์จากข้อมูลทั้งหมด แต่ไม่ได้แจกแจงรายละเอียดของความผิดพลาด ในทางการแพทย์ ความผิดพลาดบางประเภทส่งผลกระทบร้ายแรงกว่าประเภทอื่น**

**ปัญหานี้มักเกิดขึ้นในชุดข้อมูลที่ไม่สมดุล (Imbalanced Data) ซึ่งเป็นเรื่องปกติในการตรวจโรค (จำนวนผู้ป่วยมักน้อยกว่าคนปกติ) แบบจำลองอาจมี Accuracy สูงได้โดยการทายผลส่วนใหญ่เป็นคลาสที่มีจำนวนเยอะที่สุด (เช่น "เนื้อดี") ในขณะที่อาจทายผลคลาสที่สำคัญที่สุดแต่มีจำนวนน้อย ("เนื้อร้าย") ผิดพลาดเกือบทั้งหมด**

**(2) ระหว่าง False Positive (FP) และ False Negative (FN) แบบใดส่งผลร้ายแรงกว่ากัน**

**ในบริบททางการแพทย์เพื่อคัดกรองเนื้อร้าย False Negative (FN) ส่งผลกระทบร้ายแรงกว่า อย่างชัดเจน**

* **False Negative (FN): คือการที่แบบจำลองทายว่า "เนื้อดี" (Negative) ทั้งที่ความจริงแล้วเป็น "เนื้อร้าย" (Positive)**
  + **ผลกระทบ: ผู้ป่วยที่เป็นมะเร็งจะไม่ได้รับการวินิจฉัยและการรักษาที่ทันท่วงที ทำให้โรคลุกลามและอาจเป็นอันตรายถึงชีวิตได้ ซึ่งเป็นความผิดพลาดที่ยอมรับไม่ได้ทางการแพทย์**
* **False Positive (FP): คือการที่แบบจำลองทายว่า "เนื้อร้าย" (Positive) ทั้งที่ความจริงแล้วเป็น "เนื้อดี" (Negative)**
  + **ผลกระทบ: ผู้ป่วยที่ไม่เป็นมะเร็งอาจถูกส่งไปตรวจเพิ่มเติม (เช่น การตัดชิ้นเนื้อไปตรวจ) ซึ่งทำให้เกิดความวิตกกังวล เสียเวลา และเสียค่าใช้จ่ายโดยไม่จำเป็น แม้จะเป็นผลเสีย แต่ก็ไม่ร้ายแรงเท่ากับการพลาดการตรวจพบมะเร็ง**

**(3) ทีมแพทย์ควรให้ความสำคัญกับ Precision หรือ Recall มากกว่ากัน**

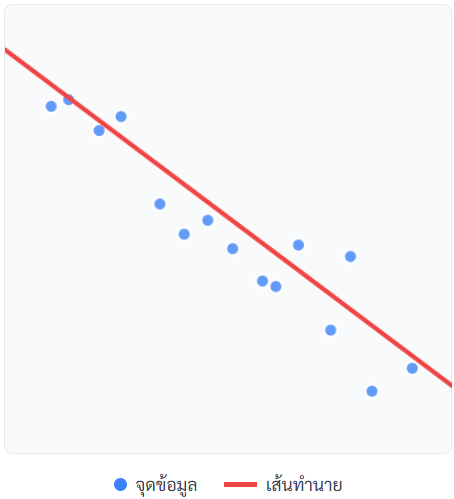
**ทีมแพทย์ควรให้ความสำคัญกับค่า Recall ของคลาส "เนื้อร้าย (Malignant)" มากที่สุด**

* **เหตุผล: Recall คือหน่วยวัดที่บอกว่า จากจำนวนเคสที่เป็น 'เนื้อร้าย' ทั้งหมด แบบจำลองสามารถตรวจจับเจอได้กี่เคส (TP / (TP + FN)) ในบริบทนี้ Recall สูงสำหรับคลาส "เนื้อร้าย" หมายความว่า แบบจำลองสามารถตรวจจับผู้ป่วยที่เป็นมะเร็งได้เกือบทั้งหมด การให้ความสำคัญกับ Recall คือการพยายามลดจำนวน False Negatives (FN) ให้เหลือน้อยที่สุด ซึ่งสอดคล้องกับเป้าหมายทางการแพทย์ที่ต้องไม่ปล่อยให้ผู้ป่วยที่เป็นโรคหลุดรอดไปโดยไม่ได้รับการวินิจฉัย**

**ข้อที่ 3: การวินิจฉัยประสิทธิภาพของแบบจำลอง (Overfitting & Underfitting)**

**(รวม 15 คะแนน | ระดับความยาก: ปานกลาง)**

**สถานการณ์:** ท่านกำลังสร้างแบบจำลองเพื่อทำนายราคาบ้านจากขนาดพื้นที่ และได้ผลลัพธ์ของแบบจำลอง 2 แบบดังภาพ:



**แบบจำลอง A: สร้างเส้นทำนายเป็นเส้นตรง (Linear Regression)**

**รูปภาพประกอบด้วย แผนภาพ, แผนที่, ไลน์, ข้อความ

เนื้อหาที่สร้างโดย AI อาจไม่ถูกต้อง**

**แบบจำลอง B: สร้างเส้นทำนายเป็นเส้นโค้งที่ซับซ้อนและพยายามวิ่งผ่านทุกจุดข้อมูล (High-degree Polynomial Regression)**

**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

**ข้อที่ 3: การวินิจฉัยประสิทธิภาพของแบบจำลอง (Overfitting & Underfitting)**

**(1) แบบจำลองใดเกิด Underfitting และ Overfitting**

* **แบบจำลอง A (เส้นตรง): มีแนวโน้มที่จะเกิดปัญหา Underfitting**
  + **เหตุผล: แบบจำลองนี้ดูเรียบง่ายเกินไป (Too Simple) และไม่สามารถจับรูปแบบหรือแนวโน้มที่แท้จริงของข้อมูลได้ดีพอ จะเห็นว่าเส้นตรงไม่สามารถปรับตัวให้เข้ากับความโค้งของกลุ่มข้อมูล ทำให้เกิดความคลาดเคลื่อนสูงทั้งกับข้อมูลที่ใช้สอนและข้อมูลใหม่**
* **แบบจำลอง B (เส้นโค้งซับซ้อน): มีแนวโน้มที่จะเกิดปัญหา Overfitting**
  + **เหตุผล: แบบจำลองนี้มีความซับซ้อนมากเกินไป (Too Complex) โดยพยายามลากเส้นให้ผ่านทุกจุดข้อมูล มันไม่ใช่แค่เรียนรู้ "แนวโน้ม" แต่เป็นการ "ท่องจำ" ข้อมูลที่ใช้สอน รวมถึงสัญญาณรบกวน (Noise) ในข้อมูลไปด้วย**

**(2) "คะแนนตอนสอน (Training Score)" และ "คะแนนตอนทดสอบ (Test Score)" ของแบบจำลอง B**

**สำหรับแบบจำลอง B ที่เกิด Overfitting:**

* **คะแนนตอนสอน (Training Score): จะมีค่า สูงมาก เข้าใกล้ 100% เพราะเส้นทำนายถูกสร้างให้วิ่งผ่านหรือเข้าใกล้จุดข้อมูลสำหรับสอนทุกจุดเท่าที่จะเป็นไปได้**
* **คะแนนตอนทดสอบ (Test Score): จะมีค่า ต่ำกว่าคะแนนตอนสอนอย่างมีนัยสำคัญ เพราะแบบจำลองที่ท่องจำ Noise มา จะไม่สามารถทำนายข้อมูลใหม่ที่ไม่เคยเห็น (Test Data) ได้ดี**

**(3) ปัญหา Underfitting และ Overfitting สะท้อนเรื่อง Bias และ Variance อย่างไร**

* **แบบจำลอง A (Underfitting): มี High Bias (ความเอนเอียงสูง) และ Low Variance**
  + **คำอธิบาย: High Bias เกิดจากแบบจำลองมีข้อสันนิษฐานที่เรียบง่ายเกินไปเกี่ยวกับข้อมูล (เช่น สันนิษฐานว่าเป็นเส้นตรง) ทำให้เกิดข้อผิดพลาดอย่างเป็นระบบและไม่สามารถเรียนรู้ความสัมพันธ์ที่แท้จริงได้**
* **แบบจำลอง B (Overfitting): มี Low Bias และ High Variance (ความแปรปรวนสูง)**
  + **คำอธิบาย: High Variance เกิดจากแบบจำลองมีความซับซ้อนและอ่อนไหวต่อข้อมูลที่ใช้สอนมากเกินไป หากเปลี่ยนชุดข้อมูลสอนเพียงเล็กน้อย เส้นทำนายที่ได้ก็จะเปลี่ยนแปลงไปอย่างมาก สะท้อนว่าแบบจำลองไม่เสถียรและไม่สามารถนำไปใช้กับข้อมูลทั่วไปได้ดี**

**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?

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

**(1) คำนวณระยะห่างแบบยูคลิด (Euclidean Distance)**

**จุดข้อมูลใหม่คือ Pnew​=(5,3). สูตรระยะห่างแบบยูคลิดคือ** 

**รูปภาพประกอบด้วย ข้อความ, ภาพหน้าจอ, ซอฟต์แวร์, ซอฟต์แวร์มัลติมีเดีย

เนื้อหาที่สร้างโดย AI อาจไม่ถูกต้อง**

**(2) ทำนายคลาสเมื่อ k = 3**

* **ขั้นตอน:**
  1. **เรียงลำดับจุดข้อมูลตามระยะห่างจากน้อยไปมาก: P4 (1.41), P3 (2.00), P2 (2.24) หรือ P6 (2.24), P5 (3.00), P1 (3.16)**
  2. **เลือกเพื่อนบ้านที่ใกล้ที่สุดมา 3 จุด (k=3) คือ: P4 (คลาส B), P3 (คลาส A), P2 (คลาส A) หรือ P6 (คลาส B)**
  3. **ลงคะแนน (Voting): คลาส A มี 2 เสียง, คลาส B มี 2 เสียง**
* **คำทำนาย: Pnew​ ควรจะอยู่ใน คลาส A หรือ คลาส B ก่อได้ !!!**

**(3) ผลการทำนายเมื่อ k = 5 และผลกระทบของการเพิ่มค่า k**

* **ผลการทำนายเมื่อ k = 5:**
  1. **เลือกเพื่อนบ้านที่ใกล้ที่สุดมา 5 จุด คือ: P4 (B), P3 (A), P2 (A), P6 (B), P5 (B)**
  2. **ลงคะแนน (Voting): คลาส B มี 3 เสียง, คลาส A มี 2 เสียง**
  3. **ผลการทำนายจะเปลี่ยนไป จากเดิมเป็นคลาส A,B จะกลายเป็น คลาส B**
* **ผลกระทบของการเพิ่มค่า k:**
  1. **การเพิ่มค่า k โดยทั่วไปจะทำให้ ขอบเขตการตัดสินใจ (Decision Boundary) ราบเรียบขึ้นและซับซ้อนน้อยลง เนื่องจากแบบจำลองจะพิจารณาเพื่อนบ้านในวงกว้างขึ้น ทำให้มีความทนทานต่อจุดข้อมูลที่เป็น Noise ได้ดีขึ้น (ลด Variance) แต่หากเพิ่ม k มากเกินไป ก็อาจทำให้แบบจำลองเรียบง่ายเกินจนเกิด Underfitting (เพิ่ม Bias) ได้**

**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?

**ข้อที่ 5: การคำนวณและวิเคราะห์เมตริกสำหรับประเมินผล (Evaluation Metrics)**

**จาก Confusion Matrix:**

* **True Positive (TP) = 70 (ทำนาย Spam ถูกว่าเป็น Spam)**
* **True Negative (TN) = 150 (ทำนาย Not Spam ถูกว่าเป็น Not Spam)**
* **False Positive (FP) = 10 (ทำนาย Spam ผิด โดยจริง ๆ เป็น Not Spam)**
* **False Negative (FN) = 20 (ทำนาย Not Spam ผิด โดยจริง ๆ เป็น Spam)**
* **Total = 150 + 10 + 20 + 70 = 250**

**(1) Accuracy**

* **สูตร: Accuracy =** 
* **วิธีทำ: Accuracy =** 
* **คำตอบ: 88%**

**(2) Precision ของคลาส "Spam"**

* **สูตร: Precision =** 
* **วิธีทำ: Precision =** 
* **คำตอบ: 87.5%**

**(3) Recall ของคลาส "Spam"**

* **สูตร: Recall =** 
* **วิธีทำ: Recall =** 
* **คำตอบ: 77.8%**

**(4) ในการคัดกรองอีเมลสแปม ระหว่าง Precision และ Recall เมตริกใดสำคัญกว่า**

**ในการคัดกรองอีเมลสแปม Precision มีความสำคัญมากกว่า**

* **เหตุผลและผลกระทบ:**
  + **Precision ต่ำ: หมายถึงมี False Positives (FP) สูง FP ในที่นี้คือ "อีเมลสำคัญ (Not Spam) ถูกทำนายผิดว่าเป็น Spam" และถูกส่งไปอยู่ในถังขยะหรือโฟลเดอร์สแปม ผลเสียคือ ผู้ใช้อาจพลาดการติดต่อสื่อสารที่สำคัญ เช่น อีเมลสมัครงาน, การแจ้งเตือนจากธนาคาร, หรือข้อความส่วนตัว ซึ่งสร้างประสบการณ์ที่แย่และเสียหายร้ายแรงกว่า**
  + **Recall ต่ำ: หมายถึงมี False Negatives (FN) สูง FN ในที่นี้คือ "อีเมลขยะ (Spam) ถูกทำนายผิดว่าเป็น Not Spam" และหลุดเข้ามาในกล่องจดหมายหลัก (Inbox) ผลเสียคือ ผู้ใช้จะรู้สึกรำคาญ ที่ต้องเห็นสแปม แต่โดยทั่วไปแล้วความเสียหายจะน้อยกว่าการพลาดอีเมลสำคัญ**

**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?

**ข้อที่ 6: การคำนวณและวิเคราะห์ Linear Regression**

**(1) คำนวณหาค่าความชัน (m) และจุดตัดแกน (c)**

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

**คำนวณค่าที่จำเป็น:**

* **n=3**
* 
* 
* 
* 

**ใช้สูตรคำนวณความชัน (m):**





**ใช้สูตรคำนวณจุดตัดแกน (c):**









**คำตอบ: ความชัน (m) ≈ 0.0605 และ จุดตัดแกน (c) ≈ 0.106**

**(2) เขียนสมการเส้นตรงและทำนายราคา**

* **สมการเส้นตรง: y= 0.0605x+0.106**
* **ทำนายราคาคอนโดขนาด 60 ตร.ม.:**
  + **แทนค่า x=60 ลงในสมการ**
  + **y = 0.0605(60)+0.106 = 3.63+0.106 = 3.736**
* **คำทำนาย: ราคาคอนโดขนาด 60 ตร.ม. คือประมาณ 3.736 ล้านบาท**

**(3) แบบจำลองนี้มีแนวโน้มที่จะมี High Bias หรือ High Variance มากกว่ากัน**

**แบบจำลอง Linear Regression ที่สร้างขึ้นนี้มีแนวโน้มที่จะมี High Bias มากกว่า**

* **เหตุผล:**
  1. **ความเรียบง่ายของแบบจำลอง: แบบจำลองเส้นตรง (Linear Regression) เป็นแบบจำลองที่เรียบง่ายและมีข้อสันนิษฐานที่ตายตัวว่าความสัมพันธ์ระหว่างขนาดและราคาเป็นเส้นตรง ซึ่งในความเป็นจริงตลาดอสังหาริมทรัพย์มีความซับซ้อนกว่านั้นมาก**
  2. **ข้อมูลน้อยเกินไป: การสร้างแบบจำลองจากจุดข้อมูลเพียง 3 จุด ไม่เพียงพอที่จะจับแนวโน้มที่แท้จริงได้ ทำให้แบบจำลองมีแนวโน้มเป็น Underfitting ซึ่งเป็นลักษณะเด่นของแบบจำลองที่มี High Bias คือ ไม่สามารถเรียนรู้ความซับซ้อนของข้อมูลได้อย่างเพียงพอ**