**เฉลยข้อสอบกลางภาค วิชา ENGCE207**

**เฉลยข้อที่ 1: ประเภทของปัญหาในการเรียนรู้ของเครื่อง (15 คะแนน)**

**Regression (การถดถอย):** คือการทำนายผลลัพธ์ที่เป็นค่าต่อเนื่อง (Continuous Value) หรือตัวเลขที่มีความหมายในเชิงปริมาณ

* **ตัวอย่าง 1:** การทำนายราคาบ้านจากขนาดพื้นที่ (ผลลัพธ์คือราคา ซึ่งเป็นตัวเลขต่อเนื่อง)
* **ตัวอย่าง 2:** การทำนายอุณหภูมิในวันพรุ่งนี้ (ผลลัพธ์คือองศาเซลเซียส ซึ่งเป็นค่าต่อเนื่อง)

**Classification (การจำแนกประเภท):** คือการทำนายผลลัพธ์ที่เป็นหมวดหมู่ (Categorical Value) หรือกลุ่มที่ไม่ต่อเนื่องกัน

* **ตัวอย่าง 1:** การคัดแยกอีเมลว่าเป็น "สแปม" หรือ "ไม่ใช่สแปม" (ผลลัพธ์มี 2 กลุ่ม)
* **ตัวอย่าง 2:** การวินิจฉัยภาพถ่ายทางการแพทย์ว่าเป็น "เนื้อร้าย" หรือ "เนื้อดี" (ผลลัพธ์มี 2 กลุ่ม)

**ความแตกต่างที่สำคัญ:** Regression ทำนาย "ปริมาณ" (How much/How many) ในขณะที่ Classification ทำนาย "ประเภท" (Which class/Which group)

**เฉลยข้อที่ 2: กระบวนการเตรียมข้อมูล (Data Preprocessing) (20 คะแนน)**

1. **จัดการรูปแบบวันที่ (order\_date):** ใช้ฟังก์ชัน pandas.to\_datetime() เพื่อแปลงคอลัมน์ order\_date ที่มีหลายรูปแบบให้เป็นข้อมูลชนิดวันที่ (datetime) ที่มีรูปแบบมาตรฐานเดียวกัน ซึ่งจะทำให้สามารถนำข้อมูลไปใช้งานต่อได้ง่าย เช่น การคำนวณหาระยะเวลา หรือการกรองข้อมูลตามช่วงเวลา
2. **จัดการค่าผิดปกติและค่าว่าง (rating):**
   * **ค่าผิดปกติ (Outliers):** สำหรับค่า 0 และ 8 ซึ่งอยู่นอกช่วง 1-5 ควรจัดการโดยอาจจะลบแถวข้อมูลนั้นทิ้ง หรือแปลงให้เป็นค่าว่าง (NaN) ก่อน
   * **ค่าว่าง (Missing Values):** หลังจากจัดการค่าผิดปกติแล้ว สามารถจัดการค่าว่างได้โดยการแทนที่ด้วยค่าที่มีความเหมาะสมทางสถิติ เช่น ค่าเฉลี่ย (Mean) หรือค่ามัธยฐาน (Median) ของคะแนนรีวิวทั้งหมด เพื่อไม่ให้ข้อมูลสูญหายไป
3. **จัดการรูปแบบราคา (price):**
   * ใช้เมธอด .str.replace() ของ Pandas เพื่อลบสัญลักษณ์สกุลเงิน (฿) และเครื่องหมายจุลภาค (,) ออกจากข้อความ
   * จากนั้นใช้เมธอด .astype(float) เพื่อแปลงชนิดข้อมูลจากข้อความ (string) ให้เป็นตัวเลขทศนิยม (float) เพื่อให้สามารถนำไปคำนวณทางคณิตศาสตร์ได้

**เฉลยข้อที่ 3: การเลือกใช้แผนภูมิเพื่อการนำเสนอข้อมูล (20 คะแนน)**

1. **ดูการกระจายตัวของเงินเดือน:** ใช้ **ฮิสโทแกรม (Histogram)** หรือ **Kernel Density Estimation (KDE) Plot** เหตุผล: แผนภูมิทั้งสองประเภทนี้สามารถแสดงให้เห็นถึงการกระจายตัวของข้อมูลตัวเลขหนึ่งตัวแปรได้ดี ทำให้เห็นว่าพนักงานส่วนใหญ่มีเงินเดือนอยู่ในช่วงใด มีการกระจุกตัวหรือเบ้ไปทางใด
2. **เปรียบเทียบเงินเดือนเฉลี่ยระหว่างแผนก:** ใช้ **แผนภูมิแท่ง (Bar Plot)** หรือ **แผนภูมิกล่อง (Box Plot)**
   * **Bar Plot:** เหมาะสำหรับการเปรียบเทียบค่าเฉลี่ยของแต่ละแผนกอย่างชัดเจน
   * **Box Plot:** ให้ข้อมูลที่ละเอียดกว่า โดยนอกจากจะเปรียบเทียบค่ามัธยฐาน (ใกล้เคียงค่าเฉลี่ย) แล้ว ยังแสดงการกระจายตัว, ควอร์ไทล์, และค่าผิดปกติ (Outliers) ของเงินเดือนในแต่ละแผนกได้ด้วย
3. **สำรวจความสัมพันธ์ระหว่างเงินเดือนและอายุงาน:** ใช้ **แผนภูมิกระจาย (Scatter Plot)** เหตุผล: แผนภูมินี้เหมาะสมที่สุดสำหรับการแสดงความสัมพันธ์ระหว่างข้อมูลตัวเลขสองตัวแปร (เงินเดือนและอายุงาน) โดยแต่ละจุดบนแผนภูมิจะแทนพนักงานหนึ่งคน ทำให้มองเห็นแนวโน้มความสัมพันธ์ได้ เช่น อายุงานมากขึ้น เงินเดือนมีแนวโน้มสูงขึ้นหรือไม่

**เฉลยข้อที่ 4: การตีความแบบจำลอง Linear Regression (15 คะแนน)**

1. **ความหมายของ Slope (4.5):** หมายความว่า **ในทุกๆ 1 ชั่วโมงที่นักศึกษานอนหลับเพิ่มขึ้นในคืนก่อนสอบ คาดว่าคะแนนสอบปลายภาคจะเพิ่มขึ้นโดยเฉลี่ย 4.5 คะแนน**
2. **ความหมายของ Intercept (50):** เป็นค่าทางทฤษฎี หมายความว่า **หากนักศึกษาไม่ได้นอนหลับเลยในคืนก่อนสอบ (X=0) แบบจำลองจะทำนายว่านักศึกษาคนนั้นจะได้คะแนนสอบเริ่มต้นที่ 50 คะแนน**
3. **ข้อจำกัดหรือสมมติฐาน:**
   * **ความสัมพันธ์เป็นเส้นตรง:** แบบจำลองสมมติว่าความสัมพันธ์ระหว่างชั่วโมงการนอนและคะแนนสอบเป็นเส้นตรง ซึ่งในความเป็นจริงอาจไม่ใช่ (เช่น การนอนมากเกินไปอาจไม่ช่วยให้คะแนนเพิ่มขึ้นในอัตราเดิม)
   * **ปัจจัยอื่นๆ:** แบบจำลองนี้ไม่ได้พิจารณาปัจจัยอื่นที่ส่งผลต่อคะแนนสอบ เช่น เวลาที่ใช้อ่านหนังสือ, ความถนัดในวิชา, หรือสุขภาพ

**เฉลยข้อที่ 5: การคำนวณ Linear Regression (25 คะแนน)**

**1. คำนวณ m และ c:**

* ข้อมูล: (2, 1.0), (4, 2.5), (6, 3.5)
* คำนวณค่าเฉลี่ย:
  + 
  + 
* คำนวณความชัน (m):
  + 
  + ตัวเศษ :
    - 
    - 
    - 
  + ตัวส่วน :
    - 
  + 
* คำนวณจุดตัดแกน (c):
  + 
  + 

**2. เขียนสมการและทำนายผล:**

* **สมการ:** 
* **ทำนายเมื่อ x = 5:**
  + 
  + คำตอบ: คาดว่าจะลดน้ำหนักได้ประมาณ 2.96 กิโลกรัม

**เฉลยข้อที่ 6: การคำนวณ k-Nearest Neighbors (k-NN) (25 คะแนน)**

**รูปภาพประกอบด้วย ข้อความ, ตัวอักษร, ภาพหน้าจอ, จำนวน

เนื้อหาที่สร้างโดย AI อาจไม่ถูกต้อง**

**2. ทำนายเมื่อ k = 3:**

* เพื่อนบ้านที่ใกล้ที่สุด 3 อันดับแรกคือ:
  1. P2 (ระยะทาง 1.41) - คลาส B
  2. P1 (ระยะทาง 2.24) - คลาส A
  3. P3 และ P6 (ระยะทาง 2.83) - คลาส B และ A (เลือกมา 1 จุด เช่น P3)
* การโหวต: มีคลาส B 1 เสียง และคลาส A 1 เสียง และ คลาส B 1 เสียง (จาก P2, P1, P3) -> คลาส B ชนะ 2:1
* คำตอบ: P\_new ควรอยู่ใน คลาส B

**3. ทำนายเมื่อ k = 1:**

* เพื่อนบ้านที่ใกล้ที่สุด 1 อันดับแรกคือ P2 (ระยะทาง 1.41) ซึ่งอยู่ในคลาส B
* ผลลัพธ์จะเหมือนกับข้อ 2
* ผลกระทบของ k ที่เล็กเกินไป: การใช้ k=1 ทำให้แบบจำลองมีความอ่อนไหวต่อข้อมูลรบกวน (Noise) หรือค่าผิดปกติ (Outliers) มากเกินไป หากจุดที่ใกล้ที่สุดบังเอิญเป็น Noise แบบจำลองก็จะทำนายผิดพลาดได้ง่าย ทำให้ความเสถียรของแบบจำลองต่ำ

**เฉลยข้อที่ 7: Overfitting, Underfitting, Bias และ Variance (20 คะแนน)**

1. **Overfitting vs Underfitting:**
   * Underfitting (เรียนรู้ได้ไม่ดีพอ): คือภาวะที่แบบจำลองเรียบง่ายเกินไป ไม่สามารถจับรูปแบบที่แท้จริงของข้อมูลได้ ทำให้มีประสิทธิภาพต่ำทั้งกับข้อมูลที่ใช้สอน (Training Data) และข้อมูลใหม่ (New Data)
   * Overfitting (เรียนรู้ได้ดีเกินไป): คือภาวะที่แบบจำลองซับซ้อนเกินไป จนไปเรียนรู้แม้กระทั่ง Noise หรือรายละเอียดปลีกย่อยของข้อมูลที่ใช้สอน ทำให้มีประสิทธิภาพสูงมากกับข้อมูลสอน แต่กลับมีประสิทธิภาพต่ำเมื่อนำไปใช้กับข้อมูลใหม่
2. **ความสัมพันธ์กับ Bias และ Variance:**
   * Underfitting สัมพันธ์กับ High Bias (ความเอนเอียงสูง): Bias คือความผิดพลาดที่เกิดจากสมมติฐานที่ไม่ถูกต้องของแบบจำลอง แบบจำลองที่เรียบง่ายเกินไป (High Bias) จะมีสมมติฐานที่ห่างไกลจากความเป็นจริง ทำให้ไม่สามารถเรียนรู้ความสัมพันธ์ที่ซับซ้อนได้
   * Overfitting สัมพันธ์กับ High Variance (ความแปรปรวนสูง): Variance คือความอ่อนไหวของแบบจำลองต่อการเปลี่ยนแปลงของข้อมูลที่ใช้สอน แบบจำลองที่ซับซ้อนเกินไป (High Variance) จะเปลี่ยนแปลงไปอย่างมากเมื่อข้อมูลสอนเปลี่ยนไปเพียงเล็กน้อย ทำให้ผลการทำนายไม่เสถียร

**เฉลยข้อที่ 8: การคำนวณเมตริกสำหรับประเมินผล (20 คะแนน)**

* **TP = 80, TN = 400, FP = 20, FN = 30**
* **Total = 80 + 400 + 20 + 30 = 530**

1. **Accuracy:**
   * **สูตร: (TP + TN) / Total**
   * **คำนวณ: (80 + 400) / 530 = 480 / 530 ≈ 0.906**
2. **Precision (ของคลาส "มีตำหนิ"):**
   * **สูตร: TP / (TP + FP)**
   * **คำนวณ: 80 / (80 + 20) = 80 / 100 = 0.8**
3. **Recall (ของคลาส "มีตำหนิ"):**
   * **สูตร: TP / (TP + FN)**
   * **คำนวณ: 80 / (80 + 30) = 80 / 110 ≈ 0.727**
4. **F1-Score (ของคลาส "มีตำหนิ"):**
   * **สูตร: 2 \* (Precision \* Recall) / (Precision + Recall)**
   * **คำนวณ: 2 \* (0.8 \* 0.727) / (0.8 + 0.727) = 2 \* 0.5816 / 1.527 ≈ 0.762**

**เฉลยข้อที่ 9: การประเมินผลแบบจำลองในบริบทที่สำคัญ (20 คะแนน)**

1. **เหตุผลที่ Accuracy ทำให้เข้าใจผิด:** ในกรณีที่ข้อมูลไม่สมดุลอย่างรุนแรง (เช่น ธุรกรรมฟอกเงินมีเพียง 0.5% จากทั้งหมด) แบบจำลองที่ทายว่า "ปกติ" ทั้งหมด ก็อาจมีค่า Accuracy สูงถึง 99.5% ได้ แต่ในความเป็นจริงแล้วแบบจำลองนั้นไม่มีความสามารถในการตรวจจับธุรกรรมที่ฟอกเงินได้เลย ดังนั้น Accuracy จึงไม่ใช่ตัวชี้วัดที่ดีในกรณีนี้
2. **ผลกระทบของ FP vs FN:**
   * False Positive (FP): ทายว่าธุรกรรมปกติเป็นการฟอกเงิน ส่งผลให้ลูกค้าผู้บริสุทธิ์ถูกตรวจสอบหรืออายัดบัญชี ทำให้เกิดความไม่พอใจและเสียประสบการณ์ แต่ธนาคารไม่เสียหายทางการเงิน
   * False Negative (FN): ทายว่าธุรกรรมฟอกเงินเป็นธุรกรรมปกติ ส่งผลกระทบร้ายแรงกว่ามาก เพราะทำให้ธนาคารไม่สามารถตรวจจับอาชญากรรมได้ ซึ่งอาจนำไปสู่ความเสียหายทางการเงินมหาศาล, การถูกปรับจากหน่วยงานกำกับดูแล, และการเสียชื่อเสียงของธนาคาร
3. **ควรให้ความสำคัญกับ Precision หรือ Recall:**
   * ในบริบทนี้ควรให้ความสำคัญกับ Recall มากที่สุด
   * เหตุผล: Recall คือความสามารถของแบบจำลองในการตรวจจับเคสที่เป็น Positive (ธุรกรรมฟอกเงิน) ได้ทั้งหมด การมี Recall สูงหมายความว่าแบบจำลองจะสามารถค้นหาและแจ้งเตือนธุรกรรมที่น่าสงสัยได้เกือบทั้งหมด ลดโอกาสที่อาชญากรรมจะหลุดรอดการตรวจสอบไปได้ แม้ว่าอาจจะต้องแลกมาด้วยการมี False Positive เพิ่มขึ้นบ้าง (Precision ต่ำลง) แต่ความเสียหายจากการตรวจสอบลูกค้าผู้บริสุทธิ์นั้นน้อยกว่าความเสียหายจากการปล่อยให้การฟอกเงินเกิดขึ้น

**เฉลยข้อที่ 10: ภาพรวมกระบวนการสร้างแบบจำลอง (20 คะแนน)**

**กระบวนการสร้างแบบจำลองเพื่อทำนายการยกเลิกบริการ (Churn Prediction) มีขั้นตอนหลักดังนี้:**

1. **การเตรียมข้อมูล (Data Preparation):**
   * ขั้นตอน: รวบรวมและทำความสะอาดข้อมูลดิบ เช่น ระยะเวลาที่ใช้บริการ, ค่าบริการรายเดือน, จำนวนครั้งที่ติดต่อฝ่ายบริการลูกค้า
   * ความท้าทาย/สิ่งที่ต้องพิจารณา: การจัดการกับข้อมูลที่ขาดหาย (Missing Values) เช่น ลูกค้าบางรายอาจไม่มีข้อมูลการติดต่อ จะต้องตัดสินใจว่าจะลบแถวนั้นทิ้ง หรือจะแทนที่ด้วยค่าที่เหมาะสม เช่น 0
2. **การสำรวจข้อมูลเชิงวิเคราะห์ (Exploratory Data Analysis - EDA):**
   * ขั้นตอน: สร้างภาพข้อมูลเพื่อทำความเข้าใจลักษณะและความสัมพันธ์ของข้อมูล เช่น สร้าง Box Plot เพื่อเปรียบเทียบค่าบริการรายเดือนระหว่างกลุ่มที่ Churn และไม่ Churn
   * ความท้าทาย/สิ่งที่ต้องพิจารณา: การตรวจจับความไม่สมดุลของข้อมูล (Imbalanced Data) ซึ่งเป็นเรื่องปกติในปัญหา Churn (คนไม่ Churn มักจะเยอะกว่า) ซึ่งอาจต้องใช้เทคนิคการสุ่มตัวอย่าง (Sampling) เช่น SMOTE เพื่อแก้ปัญหานี้ในภายหลัง
3. **การสร้างและคัดเลือกคุณลักษณะ (Feature Engineering & Selection):**
   * ขั้นตอน: สร้าง Feature ใหม่ๆ ที่อาจมีประโยชน์ เช่น "ค่าบริการต่อปี" หรือ "อัตราการติดต่อฝ่ายบริการ" และคัดเลือกเฉพาะ Feature ที่ส่งผลต่อการ Churn
   * ความท้าทาย/สิ่งที่ต้องพิจารณา: การหลีกเลี่ยงการรั่วไหลของข้อมูล (Data Leakage) ต้องแน่ใจว่า Feature ที่สร้างขึ้นไม่ได้นำข้อมูลจากอนาคตมาใช้ เช่น การใช้ข้อมูลเกี่ยวกับโปรโมชันที่ลูกค้าได้รับ *หลังจาก* ที่ตัดสินใจ Churn ไปแล้ว
4. **การสร้างและฝึกสอนแบบจำลอง (Model Building & Training):**
   * ขั้นตอน: แบ่งข้อมูลเป็นชุดสำหรับสอน (Training Set) และทดสอบ (Testing Set) แล้วเลือกอัลกอริทึม Classification มาสร้างแบบจำลอง
   * ความท้าทาย/สิ่งที่ต้องพิจารณา: การเปรียบเทียบประสิทธิภาพของอัลกอริทึมต่างๆ ควรทดลองสร้างแบบจำลองจากหลายๆ อัลกอริทึม (เช่น k-NN, Logistic Regression) และปรับจูน Hyperparameter เพื่อหาแบบจำลองที่ดีที่สุด
5. **การประเมินผลแบบจำลอง (Model Evaluation):**
   * ขั้นตอน: นำแบบจำลองมาทดสอบกับ Testing Set เพื่อวัดประสิทธิภาพ
   * ความท้าทาย/สิ่งที่ต้องพิจารณา: การเลือกใช้เมตริกวัดผลที่เหมาะสม เนื่องจากข้อมูลไม่สมดุล ควรให้ความสำคัญกับ Recall และ F1-Score ของคลาส "Churn" มากกว่า Accuracy เพราะเป้าหมายคือการค้นหาลูกค้าที่กำลังจะยกเลิกบริการให้ได้มากที่สุด เพื่อหาทางรักษาลูกค้าไว้