**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

**คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา**

**ข้อสอบกลางภาค ภาคการศึกษาที่ 1/2568**

**รายวิชา:** ENGCE207 Advanced Topics in Computer Engineering

**เวลาสอบ:** 180 นาที

**คะแนนเต็ม:** 200 คะแนน

**คำชี้แจง:**

1. ข้อสอบมีทั้งหมด 10 ข้อ แต่ละข้อมี 20 คะแนน
2. โปรดเขียนคำตอบลงในกระดาษคำตอบที่จัดเตรียมไว้ให้
3. ไม่อนุญาตให้นำอุปกรณ์สื่อสารทุกชนิดเข้าห้องสอบ
4. อนุญาตให้ใช้เครื่องคิดเลขพื้นฐานได้
5. โปรดเขียนชื่อ-นามสกุล และรหัสนักศึกษาให้ชัดเจน

**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

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

**คำถาม:** จงอธิบายความแตกต่างที่สำคัญระหว่างปัญหาแบบ **"Regression"** และ **"Classification"** ในการเรียนรู้ของเครื่อง (Machine Learning) คืออะไร? พร้อมยกตัวอย่างของปัญหาแต่ละประเภทมาอย่างน้อยประเภทละ 2 ตัวอย่างที่แตกต่างกัน

**เกณฑ์การให้คะแนน (15 คะแนน):**

* **(7 คะแนน)** อธิบายความแตกต่างระหว่าง Regression (การทำนายค่าต่อเนื่อง) และ Classification (การทำนายประเภทหรือหมวดหมู่) ได้อย่างถูกต้องและชัดเจน
* ***(8 คะแนน)*** ยกตัวอย่างปัญหาของแต่ละประเภทได้ถูกต้องและเหมาะสม ประเภทละ 2 ตัวอย่าง (ตัวอย่างละ 2 คะแนน)

**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

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

**สถานการณ์:** ท่านได้รับชุดข้อมูลการสั่งซื้อสินค้าออนไลน์ซึ่งมีปัญหาคุณภาพของข้อมูลดังต่อไปนี้:

1. คอลัมน์ order\_date มีการเก็บข้อมูลวันที่ในหลายรูปแบบปะปนกัน เช่น "2023-08-01", "02/08/2023", และ "Aug 3, 2023"
2. คอลัมน์ rating ซึ่งควรจะมีค่าคะแนนรีวิวระหว่าง 1 ถึง 5 แต่กลับพบค่าที่ผิดปกติ เช่น 0, 8, และค่าว่าง (Missing Value)
3. คอลัมน์ price ถูกเก็บในรูปแบบข้อความ (string) และมีสัญลักษณ์สกุลเงินกับเครื่องหมายจุลภาคปนอยู่ เช่น "฿2,450.75"

**คำถาม:** จงอธิบายขั้นตอนการจัดการข้อมูล (Data Cleaning & Preprocessing) ทั้ง 3 ปัญหาข้างต้นตามหลักการที่ได้เรียนมาในสัปดาห์ที่ 3 (Pandas) เพื่อเตรียมข้อมูลให้พร้อมสำหรับการนำไปวิเคราะห์ต่อไป

**เกณฑ์การให้คะแนน (20 คะแนน):**

* **(7 คะแนน)** อธิบายวิธีจัดการรูปแบบวันที่ที่หลากหลาย (เช่น การใช้ pd.to\_datetime เพื่อแปลงให้เป็นรูปแบบมาตรฐาน) ได้ถูกต้อง
* **(7 คะแนน)** อธิบายวิธีจัดการค่าผิดปกติ (Outliers) และค่าว่าง (Missing Values) ในคอลัมน์ rating ได้อย่างเหมาะสม (เช่น การลบแถว, การแทนที่ด้วยค่าเฉลี่ย/มัธยฐาน)
* **(6 คะแนน)** อธิบายวิธีแปลงคอลัมน์ price จากข้อความเป็นตัวเลข (เช่น การลบสัญลักษณ์และเครื่องหมายจุลภาค แล้วแปลงชนิดข้อมูล) ได้ถูกต้อง

**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

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

**สถานการณ์:** ท่านกำลังวิเคราะห์ชุดข้อมูลพนักงานของบริษัทแห่งหนึ่ง ซึ่งประกอบด้วยคอลัมน์ Department (แผนก), Salary (เงินเดือน), และ Years\_at\_Company (อายุงาน)

**คำถาม:** จงเลือกประเภทของแผนภูมิ (Chart) จากไลบรารี Matplotlib หรือ Seaborn ที่เหมาะสมที่สุดสำหรับตอบคำถามทางธุรกิจต่อไปนี้ พร้อมอธิบายเหตุผลประกอบ

1. เพื่อต้องการดูการกระจายตัวของเงินเดือน (Salary) ของพนักงานทั้งหมดในบริษัท
2. เพื่อต้องการเปรียบเทียบเงินเดือนเฉลี่ย (Salary) ระหว่างแผนก (Department) ต่างๆ
3. เพื่อต้องการสำรวจความสัมพันธ์ระหว่างเงินเดือน (Salary) และอายุงาน (Years\_at\_Company)

**เกณฑ์การให้คะแนน (20 คะแนน):**

* **(5 คะแนน)** เลือกแผนภูมิและให้เหตุผลสำหรับการดูการกระจายตัวของเงินเดือนได้ถูกต้อง (เช่น Histogram, KDE Plot)
* **(5 คะแนน)** เลือกแผนภูมิและให้เหตุผลสำหรับการเปรียบเทียบเงินเดือนเฉลี่ยระหว่างแผนกได้ถูกต้อง (เช่น Bar Plot, Box Plot)
* **(10 คะแนน)** เลือกแผนภูมิและให้เหตุผลสำหรับการสำรวจความสัมพันธ์ระหว่างเงินเดือนและอายุงานได้ถูกต้อง (เช่น Scatter Plot)

**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

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

**สถานการณ์:** ท่านได้สร้างแบบจำลอง Linear Regression เพื่อทำนาย "คะแนนสอบปลายภาค" (Y) ของนักศึกษาจาก "จำนวนชั่วโมงที่นอนหลับในคืนก่อนสอบ" (X) และได้สมการเส้นตรงคือ



**คำถาม:**

1. จงอธิบายความหมายของค่าความชัน (Slope) ที่เท่ากับ 4.5 ในบริบทของปัญหานี้
2. จงอธิบายความหมายของค่าจุดตัดแกน (Intercept) ที่เท่ากับ 50 ในบริบทของปัญหานี้
3. ท่านคิดว่าแบบจำลองนี้มีข้อจำกัดหรือสมมติฐานใดที่อาจไม่เป็นจริงในสถานการณ์จริง?

**เกณฑ์การให้คะแนน (15 คะแนน):**

* **(5 คะแนน)** อธิบายความหมายของ Slope (4.5) ในบริบทของปัญหาได้ถูกต้อง (เช่น ทุกๆ ชั่วโมงการนอนที่เพิ่มขึ้น จะทำให้คะแนนสอบเพิ่มขึ้นโดยเฉลี่ย 4.5 คะแนน)
* **(5 คะแนน)** อธิบายความหมายของ Intercept (50) ในบริบทของปัญหาได้ถูกต้อง (เช่น หากนักศึกษาไม่ได้นอนเลย (X=0) คาดว่าจะมีคะแนนสอบเริ่มต้นที่ 50 คะแนน)
* **(5 คะแนน)** อธิบายข้อจำกัดหรือสมมติฐานของแบบจำลองได้อย่างสมเหตุสมผล (เช่น ความสัมพันธ์อาจไม่ใช่เส้นตรงเสมอไป, มีปัจจัยอื่นที่ส่งผลต่อคะแนน)

**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

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

**สถานการณ์:** กำหนดให้มีชุดข้อมูลจำนวนชั่วโมงที่ใช้ในการออกกำลังกายต่อสัปดาห์และน้ำหนักที่ลดลงในหนึ่งเดือน ดังนี้:

|  |  |  |  |
| --- | --- | --- | --- |
| **ชั่วโมงออกกำลังกาย (X)** | 2 | 4 | 6 |
| **น้ำหนักที่ลด (กก.), Y** | 1.0 | 2.5 | 3.5 |

**คำสั่ง:**

1. จากข้อมูลที่กำหนดให้ จงคำนวณหาค่าความชัน (m) และจุดตัดแกน (c) ของสมการเส้นตรง  (แสดงวิธีทำ)
2. จงเขียนสมการเส้นตรงที่ได้ และใช้สมการดังกล่าวทำนายน้ำหนักที่จะลดลง หากออกกำลังกาย 5 ชั่วโมงต่อสัปดาห์

**เกณฑ์การให้คะแนน (25 คะแนน):**

* **(15 คะแนน)** แสดงขั้นตอนการคำนวณหาค่าความชัน (m) และจุดตัดแกน (c) ได้อย่างถูกต้องตามหลักการ
* **(10 คะแนน)** เขียนสมการเส้นตรงและใช้สมการในการทำนายค่าสำหรับ X=5 ได้อย่างถูกต้อง

**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

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

**สถานการณ์:** กำหนดให้มีชุดข้อมูลคุณสมบัติของโทรศัพท์มือถือ 2 คลาส (A: 'แนะนำ', B: 'ไม่แนะนำ') และมี **โทรศัพท์รุ่นใหม่ (P\_new)** ที่ตำแหน่ง (คุณภาพกล้อง=7, อายุแบตเตอรี่=6)

|  |  |  |  |
| --- | --- | --- | --- |
| **จุดข้อมูล** | **คุณภาพกล้อง (X)** | **อายุแบตเตอรี่ (Y)** | **คลาส** |
| **P1** | 8 | 8 | A |
| **P2** | 6 | 5 | B |
| **P3** | 9 | 4 | B |
| **P4** | 4 | 7 | A |
| **P5** | 2 | 5 | B |
| **P6** | 5 | 8 | A |

**คำสั่ง:**

1. จงคำนวณระยะห่างแบบยูคลิด (Euclidean Distance) จากจุด P\_new ไปยังจุดข้อมูลอื่นๆ ทั้ง 6 จุด (แสดงวิธีทำ)
2. จงทำนายว่า P\_new ควรจะอยู่ในคลาสใด เมื่อกำหนดให้ k=3? พร้อมอธิบายขั้นตอนการตัดสินใจ
3. หากเปลี่ยนค่า k เป็น 1 จะได้ผลลัพธ์เหมือนหรือแตกต่างจากข้อ 2 หรือไม่? และการเลือกค่า k ที่เล็กเกินไปอาจส่งผลต่อความเสถียรของแบบจำลองได้อย่างไร?

**เกณฑ์การให้คะแนน (25 คะแนน):**

* **(10 คะแนน)** คำนวณ Euclidean Distance ทั้ง 6 จุดได้ถูกต้อง (แสดงวิธีทำชัดเจน)
* **(8 คะแนน)** ทำนายคลาสเมื่อ k=3 พร้อมอธิบายขั้นตอนการหาเพื่อนบ้านที่ใกล้ที่สุด 3 อันดับแรกและการโหวตได้ถูกต้อง
* **(7 คะแนน)** ตอบคำถามและอธิบายผลกระทบของการเปลี่ยนค่า k เป็น 1 ได้อย่างสมเหตุสมผล (เช่น อาจไวต่อ Noise หรือ Outlier มากขึ้น)

**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

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

**คำถาม:**

1. จงอธิบายความหมายของปัญหา **Overfitting** และ **Underfitting** ในการสร้างแบบจำลอง Machine Learning พร้อมทั้งอธิบายว่าแต่ละปัญหาส่งผลต่อประสิทธิภาพของแบบจำลองเมื่อนำไปใช้กับข้อมูลใหม่ (New Data) อย่างไร
2. ปัญหา Overfitting และ Underfitting มีความสัมพันธ์กับ **Bias** และ **Variance** อย่างไร? จงอธิบายว่าแบบจำลองที่เกิด Overfitting มีแนวโน้มที่จะมี High Bias หรือ High Variance และแบบจำลองที่เกิด Underfitting มีแนวโน้มที่จะมี High Bias หรือ High Variance เพราะเหตุใด

**เกณฑ์การให้คะแนน (20 คะแนน):**

* **(10 คะแนน)** อธิบายความหมายและผลกระทบของ Overfitting (ทำได้ดีกับข้อมูลสอนแต่แย่กับข้อมูลทดสอบ) และ Underfitting (ทำได้แย่ทั้งกับข้อมูลสอนและทดสอบ) ได้ถูกต้อง
* **(10 คะแนน)** อธิบายความสัมพันธ์ได้ถูกต้องว่า Underfitting สัมพันธ์กับ High Bias (แบบจำลองเรียบง่ายไป) และ Overfitting สัมพันธ์กับ High Variance (แบบจำลองซับซ้อนไป) พร้อมให้เหตุผลประกอบ

**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

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

**สถานการณ์:** แบบจำลองสำหรับตรวจจับสินค้าที่ "มีตำหนิ" ในสายการผลิต ได้ผลลัพธ์เป็น Confusion Matrix ดังนี้:

|  |  |  |
| --- | --- | --- |
|  | **ทำนายว่า: ปกติ** | **ทำนายว่า: มีตำหนิ** |
| **ค่าจริง: ปกติ** | 400 (TN) | 20 (FP) |
| **ค่าจริง: มีตำหนิ** | 30 (FN) | 80 (TP) |

**คำสั่ง:** จากตาราง Confusion Matrix ข้างต้น จงคำนวณค่าต่อไปนี้สำหรับคลาส "มีตำหนิ" (แสดงสูตรและวิธีทำ):

1. Accuracy
2. Precision
3. Recall
4. F1-Score

**เกณฑ์การให้คะแนน (20 คะแนน):**

* **(5 คะแนน)** คำนวณ Accuracy ได้ถูกต้อง (แสดงสูตรและวิธีทำ)
* **(5 คะแนน)** คำนวณ Precision ได้ถูกต้อง (แสดงสูตรและวิธีทำ)
* **(5 คะแนน)** คำนวณ Recall ได้ถูกต้อง (แสดงสูตรและวิธีทำ)
* **(5 คะแนน)** คำนวณ F1-Score ได้ถูกต้อง (แสดงสูตรและวิธีทำ)

**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

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

**สถานการณ์:** ระบบธนาคารได้พัฒนาแบบจำลอง AI เพื่อตรวจจับ "ธุรกรรมที่น่าสงสัยว่าเป็นการฟอกเงิน" จากธุรกรรมปกติทั้งหมด ผลการทดสอบพบว่ามีค่า Accuracy สูงถึง 99.5% แต่ผู้ตรวจสอบภายในยังคงกังวล

**คำถาม:**

1. เหตุใดค่า Accuracy ที่สูงมาก อาจยังเป็นตัวชี้วัดที่ทำให้เข้าใจผิดได้ในบริบทที่ข้อมูลมีความไม่สมดุล (Imbalanced Data) อย่างรุนแรง เช่น การตรวจจับการฟอกเงิน
2. ในบริบทนี้ ระหว่าง **False Positive** (FP: ทายว่าธุรกรรมปกติเป็นธุรกรรมฟอกเงิน) และ **False Negative** (FN: ทายว่าธุรกรรมฟอกเงินเป็นธุรกรรมปกติ) การทายผลผิดพลาดแบบใดที่ส่งผลกระทบร้ายแรงต่อธนาคารและสังคมมากกว่ากัน? จงอธิบาย
3. เพื่อให้แบบจำลองนี้มีประโยชน์สูงสุดในการป้องกันความเสียหายทางกฎหมายและชื่อเสียง ธนาคารควรให้ความสำคัญกับการปรับปรุงค่า **Precision** หรือ **Recall** ของคลาส "ธุรกรรมฟอกเงิน" มากกว่ากัน? เพราะเหตุใด?

**เกณฑ์การให้คะแนน (20 คะแนน):**

* **(5 คะแนน)** อธิบายเหตุผลที่ Accuracy อาจไม่ใช่ตัวชี้วัดที่ดีสำหรับ Imbalanced Data ได้อย่างถูกต้อง
* **(8 คะแนน)** วิเคราะห์และเปรียบเทียบผลกระทบของ False Positive และ False Negative ในบริบทที่ให้มาได้อย่างชัดเจนและสมเหตุสมผล (โดยทั่วไป FN จะร้ายแรงกว่า)
* **(7 คะแนน)** ให้เหตุผลในการเลือกให้ความสำคัญกับ Recall ได้อย่างสมเหตุสมผล (เพื่อลดจำนวนธุรกรรมฟอกเงินที่หลุดรอดการตรวจจับให้ได้มากที่สุด)

**ชื่อ-นามสกุล:**........................................................................................ **รหัสนักศึกษา:**.............................................

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

**สถานการณ์:** ท่านได้รับมอบหมายให้สร้างแบบจำลองเพื่อทำนายว่าลูกค้าของบริษัทโทรคมนาคมมีแนวโน้มที่จะ "ยกเลิกบริการ (Churn)" หรือไม่ โดยใช้ข้อมูล เช่น ระยะเวลาที่ใช้บริการ, ค่าบริการรายเดือน, และ จำนวนครั้งที่ติดต่อฝ่ายบริการลูกค้า

**คำถาม**: จงอธิบายขั้นตอนหลักๆ ในการดำเนินโครงการนี้ ตั้งแต่การได้รับข้อมูลดิบไปจนถึงการประเมินผลแบบจำลอง Classification สุดท้าย โดยในแต่ละขั้นตอนให้ระบุถึง ความท้าทายหรือสิ่งที่ต้องพิจารณา อย่างน้อย 1 ประการ

**เกณฑ์การให้คะแนน (20 คะแนน):**

* **(10 คะแนน)** สามารถอธิบายขั้นตอนหลักๆ ได้ครบถ้วน (เช่น การเตรียมข้อมูล, การสำรวจข้อมูล, การสร้างแบบจำลอง, การประเมินผล)
* **(10 คะแนน)** สามารถระบุความท้าทายหรือสิ่งที่ต้องพิจารณาในแต่ละขั้นตอนได้อย่างสมเหตุสมผลและสอดคล้องกับหลักการ (เช่น การจัดการค่าว่าง, การเลือก Features, การเปรียบเทียบโมเดล, การเลือกเมตริกประเมินผลที่เหมาะสม)