แบบสอบกลางภาค รายวิชา 2758688 หลักการเรียนรู้ของเครื่องและการประยุกต์
ภาคต้น ปีการศึกษา 2566

คำชี้แจง

1. แบบสอบมี 2 ตอน คะแนนเต็ม 40 คะแนน ให้ทำทุกตอน
2. สอบวันที่ 6 เมษายน 2566 เวลา 08:00 น.
3. ส่งกระดาษคำตอบภายในวันที่ 7 เมษายน 2566 เวลา 08:00 น.
4. ห้ามไม่ให้มีการเผยแพร่เนื้อหาส่วนใดส่วนหนึ่งหรือทั้งหมดของแบบสอบ รวมถึงข้อมูลที่ใช้ในการสอบนี้ หากพบว่ามีพฤติกรรมดังกล่าวจะถือว่านิสิตมีพฤติกรรมที่ส่อไปในทางทุจริต
5. ให้พิมพ์ตอบและส่งกระดาษคำตอบในรูปแบบ PDF

สถานการณ์

หน่วยงานทางการศึกษาแห่งหนึ่งได้มาขอคำปรึกษากับนิสิตเกี่ยวกับการใช้ประโยชน์จากฐานข้อมูลนักเรียนของหน่วยงาน โดยฐานข้อมูลดังกล่าวประกอบด้วยข้อมูลด้านภูมิหลัง พฤติกรรรมการเรียนรู้และการดำเนินชีวิต และผลการเรียนรู้ของนักเรียนที่เรียนวิชาคณิตศาสตร์ และภาษาอังกฤษ

ตอนที่ 1 : ขอให้ใช้ชุดข้อมูล Math_miss.csv และ Eng_miss.csv เพื่อดำเนินการวิเคราะห์ ตามขั้นตอนดังนี้ (16 คะแนน)

1. รวมชุดข้อมูล Math_miss.csv และ Eng_miss.csv ให้เป็นชุดข้อมูลเดียวกัน เขียนอธิบายวิธีการดำเนินงานของนิสิตมาไม่เกิน 1 ย่อหน้า พร้อมแสดงตัวอย่างชุดข้อมูลที่รวมแล้ว (1 คะแนน)
2. สำรวจชุดข้อมูลว่ามีค่าสูญหายเกิดขึ้นหรือไม่ ค่าสูญหายที่เกิดขึ้นพบในตัวแปรใดบ้าง มีปริมาณเท่าใด เขียนอธิบายมาไม่เกิน 1 ย่อหน้า (1 คะแนน) พร้อมแสดงผลการวิเคราะห์ที่สำคัญประกอบ (1 คะแนน)
3. ตรวจสอบกลไกการสูญหายในชุดข้อมูลว่ามีแนวโน้มเป็นการสูญหายแบบใดมากกว่ากัน ระหว่าง MCAR กับ MAR โดยใช้อัลกอริทึม PCA และ logistic regression ผลการวิเคราะห์ที่ได้เป็นอย่างไร? ขอให้
    a. อธิบายวิธีการดำเนินงานวิธีการละ 1 ย่อหน้า (3 คะแนน)
    b. สรุปผลของแต่ละวิธีการ 1 ย่อหน้า (5 คะแนน)
    c. แสดงผลการวิเคราะห์ที่สำคัญประกอบ (2 คะแนน)
    d. ผลการวิเคราะห์ที่ได้ให้สารสนเทศที่เหมือนหรือแตกต่างกันอย่างไร (3 คะแนน)

การใช้ PCA วิเคราะห์กลไกค่าสูญหายอาจดำเนินการดังนี้

(1) แปลงข้อมูลของตัวแปรทุกตัวให้เป็นตัวแปรแบบให้คะแนนสอบค่า (1 และ 0) โดยที่ 1 ใช้แทนข้อมูลที่เป็นค่าสูญหาย และ 0 ใช้แทนข้อมูลที่ไม่ใช่ค่าสูญหาย

(2) นำชุดข้อมูลที่ผ่านการแปลงค่าเป็น 1, 0 ดังกล่าวมาดำเนินการวิเคราะห์ PCA (กำหนดจำนวนองค์ประกอบหลักเท่ากับ 2 องค์ประกอบ)

(3) คำนวณค่า factor score ขององค์ประกอบหลักทั้งสองตัว แล้วนำไปพล็อตแผนภาพการกระจาย

(4) ค่า factor score บนแผนภาพการกระจายสามารถใช้บอกแนวโน้มของปริมาณการสูญหายและลักษณะการสูญหายที่เกิดขึ้นในหน่วยข้อมูลได้ กล่าวคือข้อมูลใดที่มีค่า factor score อยู่ใกล้จุดกำเนิด แสดงว่าหน่วยข้อมูลนั้นมีค่าสูญหายเป็นจำนวนน้อย แต่ถ้ามีค่า factor score อยู่ไกลจุดกำเนิด แสดงว่าค่าสูญหายในหน่วยข้อมูลนั้นมีเป็นจำนวนมาก

(5) นอกจากนี้หากมีหน่วยข้อมูลที่ค่า factor score ใกล้เคียงกันเป็นกลุ่มย่อย ๆ ตั้งแต่หนึ่งกลุ่มขึ้นไป แสดงว่าการสูญหายในหน่วยข้อมูลดังกล่าวมีความสัมพันธ์กัน บ่งชี้ว่าการสูญหายในหน่วยข้อมูลดังกล่าวจะต้องมีความสัมพันธ์กับตัวแปรบางตัวที่อยู่ภายในชุดข้อมูล สถานการณ์นี้บ่งชี้ว่าการสูญหายที่เกิดขึ้นมีแนวโน้มเป็น MAR (หรือ MNAR) ในทางกลับกันหากจุดบนแผนภาพกระจายอย่างอิสระ แสดงว่าการสูญหายที่เกิดขึ้นมีแนวโน้มเป็นแบบ MCAR

ในกรณีที่เลือก decision tree อาจดำเนินการดังนี้

(1) สร้างตัวแปรใหม่แบบให้คะแนนสองค่า (1 และ 0) จากตัวแปรพบค่าสูญหาย โดยที่ 1 ใช้แทนข้อมูลที่เป็นค่าสูญหาย และ 0 ใช้แทนข้อมูลที่ไม่ใช่ค่าสูญหาย

(2) Train logistic regression โดยใช้ตัวแปรให้คะแนนสองค่าข้างต้นเป็นตัวแปรตาม และนำตัวแปรที่เหลือทั้งหมดในชุดข้อมูลเป็นตัวแปรอิสระ

(3) พิจารณาประสิทธิภาพการทำนายของโมเดลด้วยสถิติที่เหมาะสม

(4) หากประสิทธิภาพการทำนายอยู่ในระดับที่พอใช้ขึ้นไป บ่งชี้ว่าการสูญหายในตัวแปรตามที่นำมาวิเคราะห์มีแนวโน้มจะเป็นแบบ MAR

ตอนที่ 2 : สมมุติว่านิสิตดำเนินการทดแทนค่าสูญหายเรียบร้อยแล้ว ขอให้ใช้ชุดข้อมูล Math.csv และ Eng.csv ดำเนินการวิเคราะห์ข้อมูลเพื่อตอบคำถามต่อไปนี้ (24 คะแนน)

1. จากชุดข้อมูลที่กำหนดให้สามารถจัดกลุ่มนักเรียนตามตัวแปรสภาพภูมิหลัง พฤติกรรมการเรียน และผลการเรียนได้เป็นกี่กลุ่ม จากผลการแบ่งกลุ่มที่ได้ **"นักเรียนที่มีระดับผลสัมฤทธิ์ทางการเรียนแตกต่างกัน มีสภาพของภูมิหลังและพฤติกรรมการเรียน ที่แตกต่างกันอย่างไร"** เขียนอธิบายการดำเนินงานของนิสิต 1 ย่อหน้า (1 คะแนน) เขียนตอบคำถาม 1 ย่อหน้า (2 คะแนน) พร้อมแสดงผลการวิเคราะห์ที่สำคัญประกอบ (1 คะแนน)
2. **"กลุ่มนักเรียนที่มีระดับผลสัมฤทธิ์ทางการเรียนแตกต่างกัน มีสภาพของภูมิหลัง และพฤติกรรมการเรียน ที่แตกต่างกันอย่างไร"** เขียนตอบคำถาม 1 ย่อหน้า (4 คะแนน) พร้อมแสดงผลการวิเคราะห์ที่สำคัญประกอบ (1 คะแนน)
3. เลือกอัลกอริทึมมา 3 ตัว ทำการวิเคราะห์เพื่อตอบคำถามต่อไปนี้
    a. **"เป็นไปได้หรือไม่ที่จะทำนายผลสัมฤทธิ์ทางการเรียนจากตัวแปรสภาพภูมิหลัง พฤติกรรมการเรียนของนักเรียน ประสิทธิภาพในการทำนายเป็นอย่างไร"** เขียนอธิบายวิธีการดำเนินงานของนิสิต 1 ย่อหน้า (1 คะแนน) เขียนตอบคำถาม 1 ย่อหน้า (3 คะแนน) พร้อมแสดงผลการวิเคราะห์ที่สำคัญประกอบ (1 คะแนน)
    b. **"ปัจจัยใดบ้างที่มีความสำคัญหรือเป็นปัจจัยเสี่ยงต่อความสำเร็จในการเรียนของนักเรียนเมื่อสิ้นสุดภาคการศึกษา"** เขียนอธิบายวิธีการดำเนินงานของนิสิต 1 ย่อหน้า (1 คะแนน) เขียนตอบคำถาม 1 ย่อหน้า (3 คะแนน) พร้อมแสดงผลการวิเคราะห์ที่สำคัญประกอบ (1 คะแนน)
4. จากผลการวิเคราะห์ที่ได้ นิสิตมีข้อเสนอแนะในการดำเนินการเพื่อพัฒนาคุณภาพการศึกษาให้กับหน่วยงานนี้อย่างไรได้บ้าง (5 คะแนน)

เอกสารอ้างอิง

- http://www.feat.engineering/index.html
- https://www.tmwr.org/

รายละเอียดของตัวแปร

- **school** สังกัดของโรงเรียน
- **gender** เพศของนักเรียน
- **age** อายุของนักเรียน
- **location** ภูมิลำเนาของบ้านนักเรียน (U = ในเขตเทศบาล , R = นอกเขตเทศบาล)
- **famsize** จำนวนพี่น้องของนักเรียน (GT3 = มากกว่า 3 คน, LE3 = ไม่เกิน 3 คน)
- **ParentStat** สถานภาพของบิดาและมารดา (A = แยกกันอยู่ , T = อยู่ด้วยกัน)
- **MomEdu** ระดับการศึกษาของมารดา (0 = ไม่ได้เรียน, 1=ประถม, 2=ม.ต้น, 3 =ม.ปลาย, 4 = ปริญญาตรีขึ้นไป)
- **DadEdu** ระดับการศึกษาของบิดา (0 = ไม่ได้เรียน, 1=ประถม, 2=ม.ต้น, 3 =ม.ปลาย, 4 = ปริญญาตรีขึ้นไป)
- **MomJob** อาชีพมารดา
- **DadJob** อาชีพบิดา
- **StuParent** ผู้ปกครองของนักเรียน
- **traveltime** ระยะเวลาที่ใช้เดินทางจากบ้านมาโรงเรียน (1 = น้อยกว่า 15 นาที, 2 = 15 - 30 นาที, 3 = 30 นาที - 1 ชั่วโมง และ 4 = มากกว่า 1 ชั่วโมง)
- **readingtime** จำนวนชั่วโมงต่อสัปดาห์ที่นักเรียนใช้สำหรับทบทวนบทเรียน (1 = น้อยกว่า 2 ชั่วโมง, 2 = 2 - 5 ชั่วโมง, 3 = 5 - 10 ชั่วโมง และ 4 = มากกว่า 10 ชั่วโมง)
- **fail** จำนวนครั้งที่เคยสอบตกในอดีต
- **scholarship** นักเรียนได้ทุนการศึกษาหรือไม่
- **club_act** นักเรียนทำกิจกรรมชมรม/มีงานอดิเรกหรือไม่
- **nursery** ได้เรียนอนุบาลหรือไม่
- **higher** มีการวางแผนจะเรียนต่อในระดับอุดมศึกษาหรือไม่
- **internet** มีอินเทอร์เน็ตบ้านที่ใช้งานได้หรือไม่
- **InLove** มีแฟนอยู่หรือไม่
- **fam_relation** ระดับความสัมพันธ์ภายในครอบครัว (1 = แย่มาก, ..., 3 = พอใช้, ..., 5 = ดีมาก)
- **freetime** จำนวนเวลาว่างหลังเลิกเรียน (1 = น้อยมาก, ..., 3 = ปานกลาง, ..., 5 = เยอะมาก)
- **goout** ความบ่อยในการไปเที่ยวกับเพื่อน (1 = น้อยมาก, ..., 3 = ปานกลาง, ..., 5 = บ่อยมาก)
- **Drink_alc** พฤติกรรมการดื่มสุราต่อสัปดาห์ (1 = น้อยมาก/ไม่ดื่ม, 2= น้อย..., 3 = ปานกลาง, ..., 5 = บ่อยมาก)
- **health** สุขภาพโดยรวมของนักเรียน (1 = แย่มาก, ..., 3 = พอใช้, ..., 5 = ดีมาก)
- **absences** จำนวนวันที่ขาดเรียนต่อภาคการศึกษา (วัน)
- **PreTest** ความรู้พื้นฐานของนักเรียนวัดเมื่อต้นภาคการศึกษา (คะแนนเต็ม 20)
- **Ach** ผลสัมฤทธิ์ทางการเรียนเมื่อสิ้นสุดภาคการศึกษา (คะแนนเต็ม 20)