LAB Sheet: tree-based classification methods

ชื่อ-สกุล..... รหัสนักศึกษา..... ตอนเรียน .....

<u>คำสั่ง</u> อ่านชุดข้อมูล Adult Income โดยใช้ pandas พร้อมตอบคำถาม ต่อไปนี้

1. จงวิเคราะห์ข้อมูลด้วย pandas

   1.1 จำนวน label '<=50K' ..... samples และ จำนวน label '>50K' ..... samples

   1.2 features ที่เป็น categorical ได้แก่ .....

   1.3 ค่าเฉลี่ยของอายุพนักงาน เท่ากับ .....

   1.4 จำนวนพนักงานผู้ชายที่มีอายุมากกว่า 40 ขึ้นไป มีสถานะไม่เคยแต่งงาน แต่มีรายได้ (income) มากว่า 50k มีกี่คน .....

   1.5 จำนวนชั่วโมงทำงานต่อสัปดาห์ (hours-per-weeks) มากที่สุด .....

   1.6 รายการ features ที่มี missing values ('?') ได้แก่ .....

2. เขียนโค้ดจัดการกับค่า missing values ในแต่ละ feature

3. แปลง categorical features ให้อยู่ในรูปของ binary features

   จำนวน features ของชุดข้อมูลหลังทำการแปลงแล้ว .....

4. แบ่งชุดข้อมูลดังกล่าวออกเป็น 80% และ 20% สำหรับฝึก (train) และทดสอบ (test)

   จำนวน samples ในชุดข้อมูล train .....

   จำนวน samples ในชุดข้อมูล test .....

5. สร้างโมเดล Decision tree โดยใช้ชุดข้อมูล train กำหนดให้ใช้ entropy เป็น split criterion และความลึกสูงสุดของต้นไม้ไม่เกิน 2 ระดับ (level)

   5.1 รายการ features ที่ถูกใช้ใน decision tree ได้แก่.....

   5.2 feature ที่สำคัญที่สุดซึ่งจะถูกนำมาใช้พิจารณารายรับ (income) มากกว่า 50K หรือน้อยกว่า 50K คือ

   .....

   5.3 แปลงต้นไม้ตัดสินใจให้อยู่ในรูปของกฎ if-else

6. ความแม่นยำ (accuracy) ของโมเดล decision tree บนชุดข้อมูลทดสอบ .....

LAB Sheet: tree-based classification methods

ชื่อ-สกุล..... รหัสนักศึกษา..... ตอนเรียน .....

7. สร้างโมเดล random forest โดยกำหนดความลึกสูงสุดของต้นไม้เท่ากับ 3 ขณะที่จำนวนต้นไม้ย่อย (n_estimator) ใช้ GridSearchCV ค้นหาค่าที่ดีที่สุดบนชุดข้อมูล train กำหนดช่วงค่าอยู่ระหว่าง 6 ถึง 10 จำนวนต้นไม้ย่อยใน random forest ที่ให้ผลลัพธ์สูงที่สุด คือ .....

8. แสดงต้นไม้ random forest ที่สร้างได้ 6 ต้นแรก

9. ความแม่นยำ (accuracy) ของโมเดล random forest บนชุดข้อมูลทดสอบ .....

10. ปรับปรุงโมเดล random forest จากข้อ 7 โดยใช้อัลกอริทึม XgBoost พร้อมแสดงค่าความแม่นยำบนชุดข้อมูลทดสอบ (test data) .....

11. ทดลองใช้ GridSearchCV เพื่อหาจำนวนต้นไม้ย่อยใน XgBoost ที่ให้ผลลัพธ์สูงที่สุด จากการกำหนดค่าความลึกตั้งแต่ 2 ถึง 4 ระดับ และจำนวนต้นไม้ย่อยตั้งแต่ 1 ถึง 20 ต้น

12. บันทึกโมเดล XGboost ในรูปแบบของไฟล์ pickle ตั้งชื่อ XgB.pkl