Mini Project 1: การจำแนกประเภท (Classification)

สิ่งที่ต้องส่ง

1. รายงาน
   - บทนำ
     - คำอธิบายภาพรวม
     - เรื่องราวของข้อมูล การใช้ประโยชน์
   - คำอธิบายเกี่ยวกับข้อมูล
     - ลักษณะแต่ละคอลัมน์คืออะไร โดยละเอียด
   - หลักการและขั้นตอน การสกัด การเลือก และการเตรียมลักษณะ
     - อธิบายกระบวนการเตรียมข้อมูล การกำจัด Missing Value การกำจัด Noise โดยละเอียด
     - หากมีจำนวนลักษณะมาก ให้ใช้ขั้นตอนวิธี Tree-based + Permutation Importance เพื่อลดจำนวนลักษณะลง
       - ใช้ไฟล์ feature_importance.ipynb วิเคราะห์ผลลัพธ์คุณลักษณะที่มีความสำคัญ top 3 คืออะไร อธิบายว่าเกี่ยวเนื่องกับคลาสอย่างไร
       - ใช้ไฟล์ feature_selection.ipynb หาจำนวนคุณลักษณะที่เหมาะสม
     - แสดงคุณลักษณะที่เหลือหลังจากทำขั้นตอน Feature Selection
   - เทคนิคที่ใช้
     - การจัดเตรียมข้อมูลสำหรับการสร้างแบบจำลอง (Cross Validation and Evaluation sets) อาจแบ่งเป็น 80/20 หรือ 90/10 ชุดแรกใช้สร้างแบบจำลองและทวนสอบ ชุดที่สองใช้ทดสอบปัญหาการเข้ากันมากเกินไป
     - ใช้เทคนิคการจำแนกประเภทโดยอาศัยเพื่อนบ้าน (kNN), XGBoost และ เครือข่ายประสาท (Neural Network) เป็นอย่างน้อย เปรียบเทียบผล และข้อดี ข้อเสียของแต่ละวิธีเชิงการประยุกต์ใช้ขั้นตอนวิธี
     - ข้อมูล Cross Validation Set ให้ทำการค้นหาค่า Hyperparameter ที่ดีที่สุดจากเซตของค่าที่เป็นไปได้ โดยวิธี Grid search และแสดงค่า Best Parameter Set ของแต่ละเทคนิค
   - การประเมินประสิทธิภาพของแบบจำลอง
     - วิเคราะห์ผล เปรียบเทียบ และอภิปรายผลโดยใช้เครื่องมือวัดประสิทธิภาพของการจำแนกประเภท แสดงผลตัวเลขประสิทธิภาพที่ได้จาก ชุดข้อมูลสำหรับทวนสอบ (Cross Validation Set) และ ชุดข้อมูลทดสอบ (Evaluation Set) คำนวณค่าความต่าง สามารถสรุปได้ว่าเกิดหรือไม่เกิดปัญหา Underfitting และ Overfitting เพราะอะไร
     - แสดงผลตัวเลขประเมินประสิทธิภาพโดยรวมจากข้อมูลทดสอบ (Evaluation Set) เปรียบเทียบและสรุปว่าควรเลือกใช้เทคนิคการจำแนกประเภทใด เพราะอะไร
2. ไฟล์ Python Code ที่อ้างถึงข้อมูลผ่าน Github หรือ แนบไฟล์ข้อมูล