# การบ้านที่ 3 (1/2564)

**รายวิชา** คพ.348 แบบจำลองสำหรับวิทยาการข้อมูล

**<u>กำหนดส่ง</u>** <u>วันที่ 13 ธันวาคม 2564 (ก่อนเวลา 23:55 น.)</u>

## คำชี้แจง/ข้อกำหนด

1. การบ้านชิ้นนี้เป็นการบ้านที่ต้องทำเป็นกลุ่ม ๆ ละ 3-4 คน นักศึกษาจะต้องทำโจทย์ปัญหาด้วยสมาชิกทุกในกลุ่มทุกข้อ หากมีการสงสัยว่ากลุ่มของนักศึกษามีการทุจริตไม่ว่าด้วยวิธีการใด ผู้สอนรายวิชานี้จะเรียกสอบสวน หากพบว่ามีการทุจริตจริง นักศึกษาจะไม่ได้รับคะแนนของการบ้านชิ้นนี้ทั้งกลุ่ม และผู้สอนขอสงวนสิทธิ์ในการดำเนินการตามระเบียบมหาวิทยาลัยจนถึงที่สุด
2. การบ้านมีทั้งหมด 3 ข้อ คะแนนเต็ม 100 คะแนน ดังนี้

| ข้อที่ | 1 | 2 | 3 (Optional) |
|---|---|---|---|
| คะแนนเต็ม | 40 | 60 | (+15) |

3. การส่งการบ้านล่าช้าเกิน 5 วัน หลังจากกำหนดส่ง ผู้สอนขอสงวนสิทธิ์ในการไม่ตรวจให้คะแนนการบ้านไม่ว่ากรณีใดทั้งสิ้น แต่หากหากนักศึกษาส่งการบ้านล่าช้าภายใน 5 วัน คะแนนของการบ้านชิ้นนี้จะถูกหักร้อยละ 20 ของคะแนนเต็มต่อ 1 วันที่ส่งล่าช้า (โดยนักศึกษามีสิทธิในการขอละเว้นการตัดคะแนนในกรณีส่งล่าช้านี้ 1 ครั้งจากจำนวนการบ้านทั้งหมด)
   * ในกรณีที่ใช้สิทธิขอส่งล่าช้าโดยไม่ถูกตัดคะแนน จะต้องใช้สิทธิของสมาชิก**<u>สองคน</u>**ในกลุ่มที่ยังไม่เคยใช้สิทธิขอส่งล่าช้า
4. การส่งการบ้านจะต้องระบุ ชื่อ-สกุล เลขทะเบียนนักศึกษาของทุกคนในกลุ่มให้ชัดเจนในไฟล์การบ้านทุกไฟล์ โดยในแต่ละไฟล์จะต้องมีรูปแบบการตั้งชื่อดังนี้ CS348_GROUP_HW03_YY_1-2564.ZZ โดยที่ YY คือข้อที่ส่ง และ ZZ คือ file extension เช่น ถ้าสมมติไฟล์ที่ส่งเป็นไฟล์ .zip สำหรับการบ้านข้อที่ 1 สามารถตั้งชื่อว่า CS348_GROUP_HW03_01_1-2564.zip นอกจากนั้นจะต้องระบุ ชื่อ-สกุล เลขทะเบียนนักศึกษาของทุกคนในกลุ่มบนไฟล์รายงานและไฟล์เขียนโปรแกรมด้วย
5. **<u>ไม่ต้องส่ง Hard copy ของการบ้านฉบับนี้</u>**
6. หากการส่งการบ้านไม่ตรงตามข้อกำหนด การบ้านชิ้นนี้อาจไม่ได้รับการตรวจให้คะแนน
7. นักศึกษามีสิทธิในการโต้แย้งคะแนนของนักศึกษาสำหรับการบ้านชิ้นนี้ภายใน 7 วันหลังจากการประกาศคะแนน หลังจากนั้นจะถือว่านักศึกษายอมรับคะแนนการบ้านชิ้นนี้โดยปราศจากข้อโต้แย้งใด ๆ

## ข้อที่ 1 (รายงาน, 40 คะแนน)

1.1 ให้ศึกษา Datasets ต่อไปนี้

A. Iris flower dataset
https://www.kaggle.com/arshid/iris-flower-dataset

B. Red Wine Quality dataset
https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

C. Loan Prediction dataset
https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset?select=train_u6lujuX_CVtuZ9i.csv

1.1.1 ระบุว่าอะไรคือ features และอะไรคือ output (target variable) สำหรับ dataset แต่ละอัน

1.1.2 ระบุว่า Features และ output แต่ละอันคืออะไร (อธิบาย) และมีชนิดเป็น discrete หรือ continuous data

1.1.3 ให้นักศึกษาเลือก 2 features จากแต่ละ Dataset และทำ Scatter plots ของ features ที่เลือก (จะต้องมีทั้งหมด 3 Scatter plots)

1.1.4 ระบุว่าจากข้อมูลในข้อ A. B. C. เป็นปัญหาที่เป็น Classification หรือ Regression พร้อมให้เหตุผลประกอบ

1.1.5 ค้นหาข้อมูลที่เป็นปัญหาอื่นๆ ที่นักศึกษาสนใจบนอินเทอร์เน็ตเพิ่มเติม 2 ปัญหา และอธิบายคร่าวๆ ว่าปัญหาคืออะไรและต้องการแก้ปัญหาอะไร โดยต้องระบุแหล่งที่มาของปัญหาด้วย (ให้ใส่ลิงก์ต้นทางของแหล่งข้อมูล) และจะต้องเป็นปัญหา Classification และ Regression อย่างละ 1 ปัญหา

## ข้อที่ 2 (รายงาน + เขียนโปรแกรม, 60 คะแนน)

2.1 ให้ออกแบบและเขียนโปรแกรมภาษา Python สำหรับสร้างโมเดลเพื่อแก้ปัญหาในข้อ *1.A 1.B. และ 1.C* โดยสามารถเลือกใช้ Classifier ที่ได้เรียนมา ตามที่นักศึกษาเห็นว่าเหมาะสมกับปัญหา (Naïve Bayes, Linear Regression, Logistic Regression) แต่จะต้องระบุปัญหาและ Classifier ที่เลือกอย่างชัดเจน

<u>* ไม่อนุญาตให้เรียกใช้ Libraries สำเร็จรูปภายนอก (แต่สามารถเรียกใช้ฟังก์ชันพื้นฐานได้)</u>

2.2 จากข้อ 2.1 ให้แยกเป็น training และ test sets ในสัดส่วนที่เหมาะสม เช่น 70:30, 80:20 เป็นต้น โดยให้ใช้ training set ในการ train โมเดล และ ใช้ test set ในการ evaluate ประสิทธิภาพของโมเดลที่เลือก โดยให้รายงานผลประสิทธิภาพของโมเดลเป็นค่าความแม่นยำ (Accuracy) สำหรับปัญหาที่เป็น Classification และ ค่า Root Mean Square Error (RMSE) สำหรับปัญหาที่เป็น Regression

2.3 ให้ระบุรายละเอียดของ Setting ต่างๆ ของโมเดลที่เขียน รวมถึง อภิปรายผลลัพธ์ที่ได้อย่างละเอียด

2.4 ให้เลือก 1 ปัญหาจาก 1.A 1.B. และ 1.C ที่เป็น Regression

2.4.1 ให้ออกแบบการทดลองเพื่อเปรียบเทียบค่า RMSE เฉลี่ย ของ Training set และ Test set สำหรับแต่ละ Iteration

2.4.2 ให้ออกแบบการทดลองเพื่อเปรียบเทียบค่าของ Squared Errors ของ Training set เมื่อมีการเปลี่ยนแปลงค่า Learning rate อภิปรายผลลัพธ์ที่ได้อย่างละเอียด

## ข้อที่ 3 (Infographic, โบนัส 15 คะแนน)

---

**ข้อนี้เป็น Optional นักศึกษาสามารถเลือกที่จะทำหรือไม่ทำก็ได้*

ให้นักศึกษาออกแบบ Infographic เพื่อนำเสนอเรื่องใดเรื่องหนึ่งที่เรียนมาต่อไปนี้

Hidden Markov Model, Naïve Bayes Classifier, Linear Regression, Logistic Regression, Random variables, Hypothesis Testing, Confidence Interval

เกณฑ์การให้คะแนน: ความถูกต้อง/ความสมบูรณ์ของเนื้อหา รูปแบบการนำเสนอชัดเจนเข้าใจได้ง่าย