**Kth Nearest Neighbor**

**ชุดข้อมูล:** Psychology.xls

**Source Code:** 6610450951\_KNN.ipynb

**Import library ที่สำคัญและอ่านข้อมูลจากชุดข้อมูล**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

ตรวจสอบคอลัมน์ว่ามีคอลัมน์ใดบ้าง และแต่ละคอลัมน์ประกอบไปด้วยค่าอะไรบ้าง

ด้วยคำสั่ง .unique สำหรับทุกคอลัมน์ที่ปรากฏ

A screenshot of a computer program

Description automatically generated

**ข้อมูลที่ได้จากชุดข้อมูลนี้**

* เพศ
  + ชาย (Male)
  + หญิง (Female)
* อายุ
  + ตัวเลขจำนวนเต็มที่อยู่ในช่วง 18 – 34 ปี
* เคยมีแฟนมาแล้ว (คน)
  + ตัวเลขจำนวนเต็มที่อยู่ในช่วง 1 – 5
  + พบค่าที่หายไป “?”
* จำนวนครั้งที่ไปออกกำลังกายต่อสัปดาห์
  + ตัวเลขจำนวนเต็มที่อยู่ในช่วง 1 – 5
  + พบค่าที่หายไป “?”
* ระยะเวลาการนอน
  + น้อยกว่า 5 ชั่วโมง
  + 5-6 ชั่วโมง
  + 7-8 ชั่วโมง
  + มากกว่า 8 ชั่วโมง
* นิสัยการกินอาหาร
  + อาหารสุขภาพ
  + อาหารทั่วไป
  + Junkfood
* เคยคิดฆ่าตัวตาย
  + Yes
  + No
* จำนวนชั่วโมงที่เข้าห้องสมุดในหนึ่งสัปดาห์
  + จำนวนเต็มที่อยู่ในช่วง 0 – 12
  + พบค่าที่หายไป “?”
* ระดับความเครียดด้านการเงิน
  + จำนวนเต็มที่อยู่ในช่วง 1 - 5
  + พบค่าที่หายไป “?”
* เคยกินยานอนหลับ
  + Yes
  + No

**และผลลัพธ์ของการเป็นโรคซึมเศร้า**

* Yes และ No

**ทดลองตัดชุดข้อมูลตัวอย่างทุกแถวที่มีสิ่งที่ไม่ทราบค่าปรากฏ**

A screen shot of a computer program

Description automatically generated

พบว่ามีชุดข้อมูลที่ถูกตัดออกไปทั้งหมด 35 ตัวอย่างจาก 502 ตัวอย่าง

**Map ชุดข้อมูลทั้งตารางให้อยู่ในรูปตัวเลข เพื่อสะดวกต่อการนำไปคิดต่อ**

**A computer screen with text on it

Description automatically generated**

กำหนดให้การ Map ข้อมูลเป็นดังนี้

* เพศ
  + ชาย -> 1
  + หญิง -> 0
* ระยะเวลาการนอน จะใช้การประมาณค่า
  + น้อยกว่า 5 ชั่วโมง -> 4.5
  + 5-6 ชั่วโมง -> 5.5
  + 7-8 ชั่วโมง -> 7.5
  + มากกว่า 8 ชั่วโมง -> 8.5
* นิสัยการกินอาหาร
  + อาหารสุขภาพ -> 0
  + อาหารทั่วไป -> 1
  + Junkfood -> 2
* เคยคิดฆ่าตัวตาย
  + Yes -> 1
  + No -> 0
* เคยกินยานอนหลับ
  + Yes -> 1
  + No - > 0

**และ Map ข้อมูลผลลัพธ์เป็น**

ภาวะซึมเศร้า

* Yes -> 1
* No -> 0

**เตรียม Split ข้อมูล Features และ Target (Training Data)**

**A screen shot of a computer code

Description automatically generated**

โดยเตรียมฟีเจอร์ X สำหรับเรียนรู้ โดยตัดคอลัมน์สุดท้าย (เฉลย) ออก

และแปลงข้อมูล (Preprocess) ด้วย Standard Scaler รวมไปถึงกำหนดให้ y คือเฉลย

จากนั้น Split ข้อมูล Train: Test ด้วยสัดส่วน 80:20 และกำหนดให้ random\_state (seed) คือ 12

**ทำ KNN 5 folds cross validation สำหรับ K ตั้งแต่ช่วง 1 ถึง 99**

A screen shot of a computer code

Description automatically generated

พบว่า ค่า K ที่ดีที่สุด คือ 60

**Visualization**

A graph on a computer screen

Description automatically generated

จากกราฟพบว่าค่า accuracy ที่สูงที่สุด อยู่ที่จุด K มีค่าเท่ากับ 60

**ดูผลลัพท์การ Predict ข้อมูล**

โดยที่คอลัมน์แรก เป็นเฉลยจากชุดข้อมูล และคอลัมน์ที่สอง เป็นสิ่งที่เกิดจากการจำแนกด้วย KNN Classifier

**A screenshot of a computer program

Description automatically generated**

**HeatMap**

**A screenshot of a computer screen

Description automatically generated**

**Classification Report**

A screenshot of a computer

Description automatically generated