การจัดกลุ่มข้อมูลแบบฟัซซี่ซีมีนจากเครื่องชั่งน้ำหนักอัจฉริยะเพื่อวิเคราะห์โรคไม่ติดต่อที่พบบ่อย

Fuzzy C-Means Clustering Based on Body Composition Scale for Analysis of Non-Communicable Diseases

ปิยธิดา นวลเหลือ

รหัสนักศึกษา 620510511

รายงานค้นคว้าอิสระชิ้นนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

สาขาคณิตศาสตร์

คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

ปีการศึกษา 2565

การจัดกลุ่มข้อมูลแบบฟัซซี่ซีมีนจากเครื่องชั่งน้ำหนักอัจฉริยะเพื่อวิเคราะห์โรคไม่ติดต่อที่พบบ่อย

(Fuzzy C-Means Clustering Based on Body Composition Scale for Analysis of Non-Communicable Diseases)

ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษา

ตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

สาขาคณิตศาสตร์

คณะกรรมการควบคุมการค้นคว้าอิสระ

.....ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ณัฐวัชร์ สนธิชัย)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ภรัณยู จันทร)

.....กรรมการ

(นายแพทย์ณัฐพล สนธิชัย)

วันที่ 18 ตุลาคม 2565

กิตติกรรมประกาศ

การศึกษาค้นคว้าอิสระนี้เป็นส่วนหนึ่งในการศึกษากระบวนวิชา 206499 (Independent Study) โดยมีวัตถุประสงค์เพื่อให้นักศึกษาสาขาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ได้ศึกษาค้นคว้าประยุกต์ใช้ทฤษฎีและแนวคิดทางคณิตศาสตร์ในการทำการค้นคว้าอิสระ

งานค้นคว้าอิสระชิ้นนี้ประสบความสำเร็จลุล่วงไปได้ด้วยดี เนื่องจากได้รับการอนุเคราะห์ช่วยเหลือจาก ผู้ช่วยศาสตราจารย์ ดร.ณัฐวัชร์ สนธิชัย ซึ่งเป็นอาจารย์ที่ให้คำปรึกษาและให้โอกาสข้าพเจ้าในการทำงานค้นคว้าอิสระชิ้นดังกล่าว และให้ความช่วยเหลือในการแนะนำ ให้ข้อมูล ความรู้ ทักษะ เทคนิคต่าง ๆ รวมถึงแนวทางการทำการค้นคว้าอิสระ อีกทั้งตรวจสอบความถูกต้องในการศึกษาอิสระนี้ให้แก่ข้าพเจ้า ด้วยความเอาใจใส่อย่างดียิ่ง จนการศึกษาค้นคว้าอิสระนี้เสร็จสมบูรณ์ได้ด้วยดี ข้าพเจ้าตระหนักถึงความตั้งใจจริงและความทุ่มเทของอาจารย์ และขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.ภรัณยู จันทร อาจารย์กรรมการสอบงานค้นคว้าอิสระในครั้งนี้ กรุณาให้ข้อเสนอแนะ ให้คำปรึกษา แก้ไขและให้แนวคิดต่าง ๆ ที่เป็นประโยชน์ในการปรับปรุงแก้ไขงานค้นคว้าอิสระฉบับนี้ให้ดียิ่งขึ้น ด้วยความเอาใจใส่อย่างดียิ่งจนการศึกษาค้นคว้าอิสระนี้เสร็จสมบูรณ์ได้ด้วยดี

ขอบพระคุณคณาจารย์ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ทุกท่านที่ได้อบรมสั่งสอนและให้ความรู้ด้านคณิตศาสตร์แก่นักศึกษา ขอขอบคุณเจ้าหน้าที่ประสานงานของทางภาควิชาคณิตศาสตร์ที่ให้คำแนะนำช่วยเหลือ บริการประสานงานและอำนวยความสะดวกต่าง ๆ เป็นอย่างดีมาโดยตลอด

ขอบพระคุณ นายแพทย์ณัฐพล สนธิชัย ผู้อำนวยการโรงพยาบาลหริภุญชัย ราม จังหวัดลำพูน กรรมการสอบงานค้นคว้าอิสระและขอขอบพระคุณ นายแพทย์ วัชระ สนธิชัย ผู้อำนวยการโรงพยาบาลหริภุญชัย เมโมเรียล ที่ให้ความอนุเคราะห์ในการเก็บรวบรวมข้อมูลค้นคว้าอิสระนี้

สุดท้ายนี้หากการศึกษาค้นคว้าเล่มนี้จะมีประโยชน์แก่ผู้สนใจข้อมูล ข้าพเจ้าขอมอบคำขอบคุณ ชื่นชมและคุณความดีให้แก่คุณพ่อ คุณแม่ ผู้ที่ให้การดูแลและเป็นกำลังใจที่สำคัญยิ่งในการทำการศึกษาค้นคว้าของข้าพเจ้า ครูอาจารย์ทุกท่านที่ได้อบรมสั่งสอนข้าพเจ้า ในส่วนของความผิดพลาดหรือบกพร่องประการใด ข้าพเจ้าขอน้อมรับผิดไว้แต่เพียงผู้เดียวและยินดีรับฟังคำแนะนำจากทุกท่านที่ได้เข้ามาศึกษาเพื่อเป็นประโยชน์ในการพัฒนาการศึกษาค้นคว้าอิสระต่อไป

ปิยธิดา นวลเหลือ

หัวข้อ (ภาษาไทย) : การจัดกลุ่มข้อมูลแบบฟัซซี่ซีมีนจากเครื่องชั่งน้ำหนักอัจฉริยะเพื่อวิเคราะห์โรคไม่ติดต่อที่พบบ่อย

(ภาษาอังกฤษ) : Fuzzy C-Means Clustering Based on Body Composition Scale for Analysis of Non-Communicable Diseases

ชื่อผู้ทำการค้นคว้าอิสระ : นางสาวปิยธิดา นวลเหลือ

รหัสประจำตัวนักศึกษา : 620510511

ชื่อประธานกรรมการควบคุมการค้นคว้าอิสระ **:** ผู้ช่วยศาสตราจารย์ ดร.ณัฐวัชร์ สนธิชัย

ชื่อกรรมการสอบ : ผู้ช่วยศาสตราจารย์ ดร.ภรัณยู จันทร

: นายแพทย์ณัฐพล สนธิชัย

บทคัดย่อ

การค้นคว้าอิสระนี้มีวัตถุประสงค์เพื่อจัดกลุ่มบุคคลกลุ่มตัวอย่าง ตามค่าที่วัดได้จากเครื่องชั่งอัจฉริยะด้วยวิธีการจัดกลุ่มแบบ Fuzzy C-Means ผ่านโปรแกรมภาษา Python เพื่อสังเกตแนวโน้มของโรคไม่ติดต่อที่พบบ่อย จากการศึกษาพบว่า ข้อมูลถูกแบ่งออกเป็น 3 กลุ่ม คือกลุ่มที่คาดการณ์ว่ามีโอกาสจะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานน้อยกว่า 30% หรือมีโอกาสที่จะเป็นโรคไขมันในเลือดสูงน้อยมาก กลุ่มที่คาดการณ์ว่ามีโอกาสเป็นโรคความดันโลหิตสูงหรือเบาหวานน้อยกว่า 30% และกลุ่มที่คาดการณ์ว่ามีโอกาสที่จะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานหรือโรคไขมันในเลือดสูงมากกว่า 30% โดยผ่านการทดสอบได้ผลลัพธ์มีความแม่นยำ 80% (8 คนต่อ 10 คน)

Abstract

The objective of this independent study is to identify the population affected by non-communicable diseases. Utilizing Fuzzy C-Means clustering and the Body Composition Scale, its measurable value is determined. Additionally, cluster sampling is utilized to observe the trend of non-communicable diseases. In addition, research revealed that data is divided into three categories: The first category, less than thirty percent of people have hypertension or diabetes. In addition, they have no chance (a slim chance) of having hyperlipidemia. In the second group, more than 30 percent of people have hypertension or diabetes, and in the third group, more than 30 percent have hypertension or diabetes or hyperlipidemia. The accuracy of the experimental procedures is 80% (8 out of 10).

สารบัญ

บทที่ 1

บทนำ (Introduction)

1.1 ความเป็นมาและความสำคัญ

ในปัจจุบันสังคมผู้สูงอายุกำลังกลายเป็นปรากฏการณ์ที่ทยอยเกิดขึ้นในหลายประเทศของกลุ่มประเทศกำลังพัฒนาตามติดกลุ่มประเทศพัฒนาแล้วที่ส่วนใหญ่ต่างก้าวสู่สังคมผู้สูงอายุไปแล้ว ในส่วนของประเทศไทยนั้นได้ก้าวสู่สังคมผู้สูงอายุ (Aging Society) แล้วตั้งแต่ประมาณปี 2548 และกำลังเข้าสู่ระดับสังคมผู้สูงอายุสมบูรณ์ (Aged Society) ในปี 2565 โดยปัจจุบันมีราว 6-8 จังหวัดของไทยได้ก้าวเป็นสังคมสูงอายุอย่างสมบูรณ์ไปแล้วก่อนหน้าจังหวัดอื่นๆ ได้แก่ ลำปาง แพร่ ชัยนาท สิงห์บุรี อ่างทอง สมุทรสงคราม และอาจรวมถึงลำพูนและอุตรดิตถ์ ซึ่งเป็นไปได้ว่าไทยจะเป็นประเทศแรกในบรรดากลุ่มประเทศกำลังพัฒนาที่เข้าสู่สังคมผู้สูงอายุสมบูรณ์ หลังจากนั้นในราวปี 2575 ไทยก็มีแนวโน้มที่จะขยับเป็นสังคมผู้สูงอายุสมบูรณ์อย่างเต็มที่ (Super-Aged Society) มีประชากรอายุ 60 ปีขึ้นไปเกินกว่า 28% ของประชากรทั้งประเทศ (ข้อมูลอ้างอิง : กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์)

และจากสถานการณ์ Covid-19 ทำให้เกิดกระแส New Normal และ Next Normal ทุกคนต้องปรับตัวเพื่อให้ดำเนินชีวิตต่อไป ทำให้ผู้คนหันมาสนใจและตระหนักถึงประเด็นการดูแลสุขภาพของตนเองและส่วนรวมมากขึ้น

จากสถานการณ์ต่าง ๆ ที่ได้กล่าวมาข้างต้น ทำให้ตระหนักถึงปัญหาสุขภาพ คุณภาพชีวิตของผู้คน เมื่อก้าวเข้าสู่วัยผู้สูงอายุ แน่นอนว่าวัยนี้จะมีการเปลี่ยนแปลงหลายด้าน ทั้งทางด้านสมอง ทางอารมณ์ และทางร่างกายที่อาจถดถอยลงไม่เหมือนตอนวัยหนุ่มสาว ซึ่งเมื่อร่างกายทำงานเสื่อมลงก็เปิดทางให้โรคต่างๆเข้ามาได้ง่าย ซึ่งโรคที่พบบ่อยในผู้สูงอายุที่ควรเฝ้าระวัง เช่น โรคเบาหวาน โรคความดันโลหิตสูง โรคอ้วน โรคไขมันในเลือดสูง โรคมะเร็ง โรคตับ โรคไต โรคหัวใจ เป็นต้น โดยโรคดังกล่าวเป็นกลุ่มโรค NCDs (Noncommunicable diseases หรือโรคไม่ติดต่อ) เป็นปัญหาสุขภาพอันดับหนึ่งของโลกทั้งในมิติของจำนวนการเสียชีวิตและภาระโรคโดยรวม จากการรายงานข้อมูลขององค์การอนามัยโลก (WHO) พบประชากรทั่วโลกเสียชีวิตจากโรค NCDs มีแนวโน้มเพิ่มขึ้นจาก 38 ล้านคน (คิดเป็นร้อยละ 68 ของสาเหตุการเสียชีวิตทั้งหมดของประชากรโลก) ในปี พ.ศ. 2555 เป็น 41 ล้านคน (คิดเป็นร้อยละ 71 ของสาเหตุการเสียชีวิตทั้งหมดของประชากรโลก) ในปี พ.ศ.2559 ซึ่งในแต่ละปีพบผู้เสียชีวิตจากโรค NCDs ในกลุ่มอายุ 30-69 ปี หรือเรียกว่า “การเสียชีวิตก่อนวัยอันควร” มากถึง 15 ล้านคน

ฉะนั้นการป้องกัน และสังเกตอาการตั้งแต่เนิ่นๆ คือสิ่งสำคัญเพื่อนำไปสู่การดูแลรักษาที่ทันท่วงที และสามารถลดความรุนแรงของโรคได้

ดังนั้นงานค้นคว้าอิสระนี้ได้เห็นความสำคัญดังกล่าว เพื่อหาวิธีการกรองผู้มีความเสี่ยงที่จะเกิดโรคไม่ติดต่อที่พบบ่อยด้วยการนำข้อมูลจากเครื่องชั่งอัจฉริยะที่สามารถหาได้ใกล้ตัว โดยศึกษาว่ามีความสัมพันธ์กับการเกิดโรคหรือไม่ และทดสอบผ่านขั้นตอนวิธีการแบ่งกลุ่มข้อมูล ในการจำแนกโรคเบื้องต้นเพื่อเฝ้าระวัง อีกทั้งเป็นข้อมูลประกอบการตัดสินใจสำหรับเข้ารับการปรึกษาทางการแพทย์ต่อไป

งานค้นคว้าอิสระได้ทำการรวบรวมข้อมูลกลุ่มตัวอย่าง ตามค่าที่วัดได้จากเครื่องชั่งอัจฉริยะ InBody Dial Body Composition Analyzer เพื่อวิเคราะห์ จัดกลุ่ม และสังเกตแนวโน้มของโรคไม่ติดต่อที่พบบ่อยจากบุคคลกลุ่มตัวอย่าง ที่ได้จากการจัดกลุ่มแบบ Fuzzy C-Means ผ่านโปรแกรมภาษา Python จากการคำนวณเห็นว่าข้อมูลถูกแบ่งออกเป็น 3 กลุ่ม คือกลุ่มที่คาดการณ์ว่ามีโอกาสจะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานน้อยกว่า 30% หรือมีโอกาสที่จะเป็นโรคไขมันในเลือดสูงน้อยมาก กลุ่มที่คาดการณ์ว่ามีโอกาสเป็นโรคความดันโลหิตสูงหรือเบาหวานน้อยกว่า 30% และกลุ่มที่คาดการณ์ว่ามีโอกาสที่จะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานหรือโรคไขมันในเลือดสูงมากกว่า 30% โดยผ่านการทดสอบได้ผลลัพธ์มีความแม่นยำ 80%

1.2 วัตถุประสงค์ของการทำการค้นคว้าอิสระ

1. เพื่อจัดกลุ่มบุคคลกลุ่มตัวอย่าง ตามค่าที่วัดได้จากเครื่องชั่งอัจฉริยะ InBody Dial Body Composition Analyzer

2. เพื่อสังเกตแนวโน้มของโรคไม่ติดต่อที่พบบ่อยจากบุคคลกลุ่มตัวอย่าง ที่ได้จากการจัดกลุ่มแบบ Fuzzy C-Means

3. เพื่อศึกษาแนวคิด วิธีการจัดกลุ่มแบบ Fuzzy C-Means

4. เพื่อศึกษาวิธีการใช้โปรแกรมภาษา Python Editor Vs Code

1.3 ประโยชน์ที่คาดว่าจะได้รับจากการค้นคว้าอิสระ

1. สามารถทราบได้ว่าแต่ละบุคคลกลุ่มตัวอย่างอยู่กลุ่มข้อมูลใด

2. สามารถวิเคราะห์โรคไม่ติดต่อที่พบบ่อยจากการจัดกลุ่มบุคคลกลุ่มตัวอย่าง

3. ผู้ศึกษาค้นคว้ามีความเข้าใจแนวคิด วิธีการจัดกลุ่มแบบ Fuzzy C-Means

4. ผู้ศึกษาค้นคว้าสามารถใช้โปรแกรมภาษา Python ได้

1.4 ขอบเขตของการค้นคว้าอิสระ

1. การศึกษาครั้งนี้ ต้องการจัดกลุ่มบุคคลกลุ่มตัวอย่าง ตามค่าที่วัดได้จากเครื่องชั่งอัจฉริยะ InBody Dial Body Composition Analyzer เพื่อสังเกตแนวโน้มของโรคไม่ติดต่อที่พบบ่อยจากบุคคลกลุ่มตัวอย่าง ที่ได้จากการจัดกลุ่มแบบ Fuzzy C-Means โดยใช้โปรแกรมภาษา Python ในการคำนวณ

2. กลุ่มตัวอย่างที่ใช้ในการศึกษาคือ บุคคลเพศชายและเพศหญิงอายุ 40 ปีขึ้นไป จำนวน 50 คน

3. ตัวแปรที่ใช้ในการศึกษาคือ

3.1 ค่าที่วัดจากเครื่องชั่งอัจฉริยะ InBody Dial Body Composition Analyzer คือ น้ำหนัก ไขมันในร่างกาย มวลกล้ามเนื้อ ไขมันในช่องท้อง

3.2 ค่าที่ใช้พิจารณาเพิ่มเติมคือ ส่วนสูง เพศ อายุ

3.3 ค่าเป้าหมายคือ โรคไม่ติดต่อที่พบบ่อย

4. ระยะเวลาที่ใช้ในการศึกษา ดำเนินการภายในภาคเรียนที่ 1 ปีการศึกษา 2565

1.5 ขั้นตอนการดำเนินการของการค้นคว้าอิสระ

1. ศึกษาแนวคิด วิธีการจัดกลุ่มแบบ Fuzzy C-Means

2. ศึกษาวิธีการใช้โปรแกรมภาษา Python

3. สุ่มบุคคลอายุ 40 ปีขึ้นไป จำนวน 50 คน มาวัดค่าร่างกายจากเครื่องชั่งอัจฉริยะ InBody Dial Body Composition Analyzer แล้วเก็บรวบรวมข้อมูล

4. นำข้อมูลมาวิเคราะห์โดยวิธีการจัดกลุ่ม Fuzzy C-Means เพื่อจัดกลุ่มบุคคล

5. วิเคราะห์โรคไม่ติดต่อที่พบบ่อยของบุคคลในแต่ละกลุ่ม

6. สรุปผลการดำเนินการ

7. จัดทำรูปเล่ม

บทที่ 2

ความรู้พื้นฐาน

2.1 Argument of the maximum and the minimum

กำหนด $f: \mathbb{R} \to \mathbb{R}$ อาร์กิวเมนต์ของค่าสูงสุด (argument of the maximum: arg max, argmax) หมายถึงค่าของอาร์กิวเมนต์ (ตัวแปรต้น) ที่ให้กับนิพจน์แล้วทำให้เกิดค่าสูงสุดบนโดเมนที่พิจารณา นั่นคือ

$$\underset{x}{\operatorname{argmax}} f(x) \in \{x | \forall y: (y \neq x \implies f(y) \leq f(x))\}$$

หรือกล่าวในอีกทางหนึ่ง

$$\underset{x}{\operatorname{argmax}} f(x)$$

คือค่าของ x ที่จะทำให้ฟังก์ชัน $f(x)$ มีค่ามากที่สุด

ตัวอย่างเช่น $f(x) = -|x|$ จะเกิดค่าสูงสุดเมื่อ $x = 0$

กรณี อาร์กิวเมนต์ของค่าสูงสุดมีสมาชิกเพียงตัวเดียว สมมติเป็น x_0 สามารถเขียนได้เป็น

$$\underset{x}{\operatorname{argmax}} f(x) = x_0$$

<u>ตัวอย่าง</u>

$$\underset{\mathrm{x}}{\operatorname{argmax}}\big(x(10-x)\big) = 5$$

ซึ่งค่าสูงสุดของ $x(10-x)$ จะเท่ากับ 25 เมื่อ $x = 5$

กรณีที่มีค่า x หลายค่าที่ทำให้เกิดค่าสูงสุด คำตอบของอาร์กิวเมนต์ของค่าสูงสุดจะเป็นเซต

$$\underset{\mathrm{x}}{\operatorname{argmax}} \cos(\mathrm{x}) = 0, 2\pi, 4\pi, \ldots$$

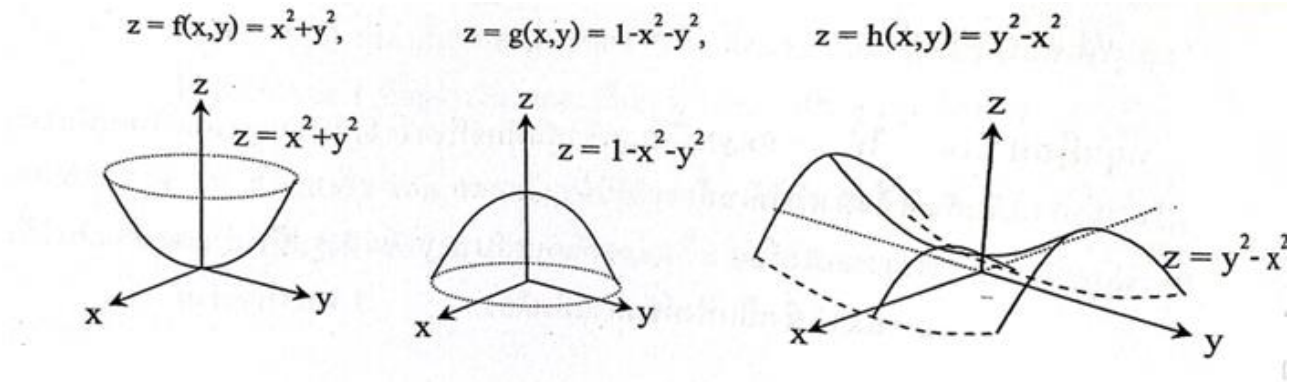
ซึ่งค่าสูงสุดของ $\cos(\mathrm{x})$ จะเท่ากับ 1 เมื่อ $x = 0, 2\pi, 4\pi, \ldots$

สำหรับ อาร์กิวเมนต์ของค่าต่ำสุด (argument of the minimum: arg min, argmin) สามารถนิยามได้ในทางตรงข้าม

2.2 จุดวิกฤตของฟังก์ชันหลายตัวแปร [1]

จุดวิกฤตของฟังก์ชัน $z = f(x,y)$ คือจุด (a,b) ที่อยู่ภายในโดเมน f โดยที่ $f_x(a,b) = 0$ และ $f_y(a,b) = 0$ หรือที่อนุพันธ์ย่อยตัวหนึ่ง หรือทั้งคู่ของ $f_x(a,b)$ และ $f_y(a,b)$ หาค่าไม่ได้

ตัวอย่าง พิจารณาฟังก์ชันต่อไปนี้



มีอนุพันธ์ย่อยดังนี้

$$f_x(x,y) = 2x, f_y(x,y) = 2y, g_x(x,y) = -2x,$$
$$g_y(x,y) = -2y, h_x(x,y) = -2x, h_y(x,y) = 2y$$

ทั้งสามฟังก์ชันมีอนุพันธ์ย่อยเทียบกับ x และเทียบกับ y เป็นศูนย์ที่จุด $(0,0)$ ทั้งหมด ดังนั้นจุด $(0,0)$ จึงเป็นจุดวิกฤตของทุกฟังก์ชัน ซึ่งจากรูปจะเห็นว่า f มีค่าต่ำสุดสัมพัทธ์ที่จุด $(0,0)$ และ g มีค่าสูงสุดสัมพัทธ์ที่จุด $(0,0)$ แต่ฟังก์ชัน h ไม่มีทั้งค่าต่ำสุดสัมพัทธ์และสูงสุดสัมพัทธ์ที่จุด $(0,0)$

2.3 เวกเตอร์เกรเดียนต์ [2]

ให้ f เป็นฟังก์ชันสเกลาร์ของสามตัวแปร x, y และ z แล้ว **เวกเตอร์เกรเดียนต์** (gradient vector) ของ f คือเวกเตอร์

$$\nabla f = \frac{\partial f}{\partial x} i + \frac{\partial f}{\partial y} j + \frac{\partial f}{\partial z} k$$

เวกเตอร์เกรเดียนต์นี้ยังสามารถเขียนแทนด้วยสัญลักษณ์ $grad\ f$ และสัญลักษณ์ ∇f นี้จะอ่านว่า “เกรเดียนต์ของ f” หรือ “เดล f” ก็ได้

ตัวอย่าง

กำหนดให้ $f(x,y,z) = xyz^2 + sin\,xy + ln\,z$ จงหา ∇f

ฟังก์ชัน f มีอนุพันธ์ย่อยเทียบกับ x คือ

$$\frac{\partial f}{\partial x} = yz^2 + y\,cos\,xy$$

ฟังก์ชัน f มีอนุพันธ์ย่อยเทียบกับ x คือ

$$\frac{\partial f}{\partial y} = xz^2 + x\,cos\,xy$$

ฟังก์ชัน f มีอนุพันธ์ย่อยเทียบกับ x คือ

$$\frac{\partial f}{\partial z} = 2xyz + \frac{1}{z}$$

ดังนั้น $\nabla f = (yz^2 + y\,cos\,xy)i + (xz^2 + x\,cos\,xy)j + (2xyz + \frac{1}{z})k$

2.4 ตัวคูณลากรานจ์ (Lagrange Multiplier) [3]

เริ่มจากปัญหาการหาค่าสูงสุดหรือต่ำสุดของฟังก์ชันสองตัวแปร $\boldsymbol{f(x,y)}$ ภายใต้เงื่อนไขบังคับ $\boldsymbol{g(x,y) = 0}$ เขียนแทนด้วย

Maximize (or minimize) $$z = f(x,y) \quad (1)$$

Subject to $$g(x,y) = 0 \quad (2)$$

กราฟของ $g(x,y) = 0$ คือเส้นโค้ง C ในระนาบ XY ดังนั้นปัญหาดังกล่าวก็คือ การหาค่าสูงสุดหรือต่ำสุดของ $\boldsymbol{f(x,y)}$ เมื่อ $\boldsymbol{(x,y)}$ แปรอยู่บนเส้นโค้งเงื่อนไข C

ถ้า (x_0, y_0) เป็นจุดบนเส้น C แล้ว จะเรียกว่า $\boldsymbol{f(x,y)}$ มี**ค่าสูงสุดสัมพัทธ์** ที่ (x_0, y_0) ถ้ามีวงกลมศูนย์กลางอยู่ที่ (x_0, y_0) โดยที่

$$f(x_0, y_0) \geq f(x,y)$$

สำหรับทุกจุด $\boldsymbol{(x,y)}$ บน C ที่อยู่ข้างในวงกลม และจะเรียกว่า $\boldsymbol{f(x,y)}$ มี**ค่าต่ำสุดสัมพัทธ์** ที่ (x_0, y_0) ถ้า

$$f(x_0, y_0) \leq f(x, y)$$

สำหรับทุกจุด (x, y) บน C ที่อยู่ข้างในวงกลม และจะเรียกว่า $f(x, y)$ มีค่าต่ำสุดสัมพัทธ์ ที่ (x_0, y_0) ถ้า

$$f(x_0, y_0) \geq f(x, y)$$

วิธีของตัวคูณลากรองจ์จะหลีกเลี่ยงการแก้สมการ (2) โดยตรง และจะสร้างฟังก์ชั่นใหม่ขึ้นมา เรียกว่า ฟังก์ชัน F โดยที่

$$F(x, y, \lambda) = f(x, y) + \lambda g(x, y)$$

เมื่อ λ (lambda อ่านว่า แลมบ์ดา) เป็นค่าคงที่เรียกว่า **ตัวคูณลากรานจ์** ค่าสูงสุดหรือต่ำสุดของ f จะเกิดที่จุด (x_0, y_0)

ค่าสูงสุดสัมพัทธ์และต่ำสุดสัมพัทธ์ของฟังก์ชัน $z = f(x, y)$ ภายใต้เงื่อนไขบังคับ $g(x, y) = 0$ จะเกิดที่จุด (x_0, y_0) ซึ่งมี (x_0, y_0, λ_0) เป็นคำตอบของระบบสมการ

$$F_x(x, y, \lambda) = 0$$
$$F_y(x, y, \lambda) = 0$$
$$F_z(x, y, \lambda) = 0$$

เมื่อ $F_x(x, y, \lambda) = f(x, y) + \lambda g(x, y)$ และอนุพันธ์ย่อยทั้งหมดของ F หาค่าได้ เรียกจุด (x_0, y_0) ว่า**จุดวิกฤติ**ของ F

ขั้นตอนวิธีของตัวคูณลากรานจ์ มีดังนี้

1. เขียนโจทย์ปัญหาให้อยู่ในรูป

Maximize (or minimize) $z = f(x, y)$

Subject to $g(x, y) = 0$

2. สร้างฟังก์ชัน F โดยที่

$$F(x, y, \lambda) = f(x, y) + \lambda g(x, y)$$

3. หาจุดวิกฤติของ F โดยการแก้ระบบสมการ

$$F_x(x, y, \lambda) = 0$$

$$F_y(x, y, \lambda) = 0$$

$$F_z(x, y, \lambda) = 0$$

4. หาค่า $z = f(x, y)$ ที่จุด (x_0, y_0) ที่มี (x_0, y_0, λ_0) สอดคล้องระบบสมการในขั้นตอน 3 จุด (x_0, y_0) อาจจะมีหลายจุด ค่าสูงสุดหรือค่าต่ำสุดของ $f(x, y)$ จะเป็นค่าใดค่าหนึ่งของค่าที่หาได้ ณ จุดวิกฤตเหล่านั้น

<u>ตัวอย่าง</u> จงหาค่าต่ำสุดของฟังก์ชัน $f(x, y) = x^2 + y^2$ ภายใต้เงื่อนไข $x + y = 10$

วิธีทำ **ขั้นที่ 1** เขียนโจทย์ปัญหาให้อยู่ในรูป

Maximize (or minimize) $z = f(x, y) = x^2 + y^2$

Subject to $g(x, y) = x + y - 0 = 0$

ขั้นที่ 2 สร้างฟังก์ชันใหม่

$$F_x(x, y, \lambda) = f(x, y) + \lambda g(x, y)$$

$$= x^2 + y^2 + \lambda(x + y - 0)$$

ขั้นที่ 3 สร้างระบบสมการ $F_x = 0$, $F_y = 0$, $F_z = 0$

$F_x = 0 \rightarrow$ $2x + \lambda = 0$ (1)

$F_y = 0 \rightarrow$ $2y + \lambda = 0$ (2)

$F_z = 0 \rightarrow$ $x + y - 10 = 0$ (3)

แก้สมการดังนี้ จาก (1) และ (2) ได้ $x = -\frac{\lambda}{2}$, $y = -\frac{\lambda}{2}$ แล้วนำไปแทนใน (3) จะได้

$$-\frac{\lambda}{2} - \frac{\lambda}{2} - 10 = 0 \rightarrow \lambda = -10$$

ดังนั้น $x = 5$, $y = 5$ จุดวิกฤติคือ $(5,5)$

ขั้นที่ 4 หาค่า f ที่จุดวิกฤต ได้

$$f(5,5) = 5^2 + 5^2 = 50$$

ตรวจสอบจุดอื่นๆ บนเส้น $x + y = 10$ ที่อยู่ใกล้ๆจุด $(5,5)$ พบว่า $f(5,5)$ ที่ได้คือค่าต่ำสุด

2.5 การวัดระยะห่างแบบยูคลิเดียน (Euclidean Distance) [4]

การวัดระยะห่างแบบยูคลิเดียน เป็นการวัดค่าความห่างระหว่างข้อมูล 2 ข้อมูล ในระบบพิกัดคาร์ทีเซียน ที่มาจากทฤษฎีปีทาโกรัส ซึ่งคำนวณได้จากสมการ ดังนี้

$$dist(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2}$$

โดยที่ $dist(x_i, x_j)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับ x_j

d คือ จำนวนมิติทั้งหมดของข้อมูล

$x_{i,k}$ คือ มิติตัวที่ k ของข้อมูล x_i

$x_{j,k}$ คือ มิติตัวที่ k ของข้อมูล x_j

ซึ่งถ้าข้อมูล 2 ตัวมีความคล้ายกันมาก แสดงว่าข้อมูลแต่ละตัวจะอยู่ใกล้กันมาก จะทำให้ค่ายูคลิเดียนมีค่าน้อย ๆ เข้าใกล้ศูนย์

2.6 Machine Learning [5]

ตามหนังสือ Deep Learning ของ Françios Chollet [3] ผู้พัฒนา Keras ได้แบ่งรูปแบบ Machine Learning ออกเป็น 4 ประเภท

2.6.1 Supervised Learning

การสร้างโมเดลเพื่อแปลงข้อมูล input เป็น target บางอย่าง ตัวอย่างง่ายที่สุดคือ classification กับ regression การจะสร้างโมเดลประเภทนี้ขึ้นมา ต้องมีชุดข้อมูลที่มีทั้ง input และ target ซึ่งจัดหามาโดยมนุษย์ เช่น การสร้าง spam filter ต้องรวมรวบข้อมูล email จำนวนมากและให้คนมาดูว่าอันไหนเป็น spam บ้าง แล้วนำมาสร้างโมเดล spam filter จากข้อมูลเหล่านี้

Supervised Learning เป็น Machine Learning ที่ถูกใช้งานมากที่สุด เข้าใจง่ายที่สุด และทุกคนที่เริ่มเรียน Machine Learning ควรเริ่มจาก Supervised Learning

2.6.2 Unsupervised Learning

การสร้างโมเดลโดยใช้ข้อมูล input เพียงอย่างเดียว ไม่มี target การใช้งานหลักมี 2 อย่างคือ

- Dimensionality reduction การลดมิติของข้อมูล เพื่อลดความซับซ้อนก่อนนำไปใช้ต่อ หรือเพื่อแสดงผลในรูปกราฟที่คนอ่านได้
- Clustering การจัดกลุ่มข้อมูลตามคุณลักษณะ เช่น การจัดกลุ่มลูกค้าตามพฤติกรรมการซื้อของ

การสร้างโมเดลประเภทนี้ขึ้นมา ใช้เพียงข้อมูล input อย่างเดียว ไม่ต้องจัดหา target เช่น โมเดลการจัดกลุ่มลูกค้า เราไม่ต้องรู้มาก่อนว่าจะมีกลุ่มอะไรบ้าง

2.6.3 Self-supervised Learning

การสร้างโมเดลด้วยวิธี Supervised Learning แต่ใช้ target แบบที่ไม่ต้องพึ่งคน เช่น

- Autoencoders ใช้ target เหมือนกับ input
- การพยากรณ์อากาศในวันถัดไปด้วยข้อมูลในอดีต ใช้ target เป็น input ในอนาคต

ปกติตำราที่แบ่ง Machine Learning เป็น 3 ประเภทจะไม่มีประเภทนี้ ในกรณีนั้นมักจะจัดรวมอยู่ใน Supervised Learning

2.6.4 Reinforcement Learning

การสอน agent ในสภาพแวดล้อมบางอย่าง ให้เรียนรู้วิธีการตัดสินใจที่ให้ผลลัพธ์ดีที่สุด ลองนึกถึงคอมพิวเตอร์ในเกมส์อะไรซักอย่างที่พยายามหาทางเอาชนะเรา

2.7 การจัดกลุ่ม (Clustering) [6]

การจัดกลุ่ม (Clustering) จัดเป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ซึ่งหมายถึงการเรียนรู้จากข้อมูลตัวอย่างที่ไม่มีการกำหนดค่าเป้าหมาย (Target) หรือฉลาก (Label) ของคลาส (Class) ไว้ ซึ่งประกอบด้วยการจัดกลุ่ม (Clustering) ด้วยวิธี K-Means, Fuzzy C-Means, Self-Organizing Map (SOM) และ Expectation-Maximization (EM) วิธีทั้ง 4 เป็นขั้นตอนวิธีที่ได้รับความนิยมอย่างสูงในการจัดกลุ่ม

การจัดกลุ่มด้วยวิธี K-Means

การจัดกลุ่มด้วยขั้นตอนวิธี K-Means เป็นขั้นตอนวิธีที่ง่ายที่สุด แต่เป็นที่นิยมในการประยุกต์ใช้งานอย่างแพร่หลายกับข้อมูลประเภทต่าง ๆ ที่มาของชื่อ K-Means เกิดจากในการเริ่มกระบวนการจัดกลุ่มนั้นจำเป็นต้องกำหนดค่า K หรือจำนวนกลุ่ม (Cluster) ก่อน การจัดกลุ่มจะทำการจัดกลุ่มของเซตตัวอย่าง $\{x_1, x_2, x_3, \dots, x_n\}$ เมื่อ $x_i \in \mathbb{R}^d$ ให้ตัวอย่างที่คล้ายกันอยู่ในกลุ่มเดียวกัน ผลลัพธ์ของการจัดกลุ่มคือกลุ่มของตัวอย่างแต่ละตัวและเซนทรอยด์ (Centroid) ซึ่งมีลักษณะเป็นตัวแทนของแต่ละกลุ่ม มีขนาดมิติเท่ากับขนาดมิติของข้อมูลตัวอย่างและมักอยู่กึ่งกลางของกลุ่มตัวอย่างในแต่ละกลุ่ม โดยสามารถสร้างฟังก์ชันวัตถุประสงค์สำหรับการจัดกลุ่มนี้ได้จากการหาค่าที่น้อยที่สุดของระยะห่างรวมของตัวอย่าง และ Centroid ของแต่ละกลุ่ม c_k ดังนี้

$$\min_{c} J(r,c) = \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\left\|x_i - c_j\right\|^2$$

โดยที่ 1. ระยะทาง $\left\|x_i - c_j\right\| = dist(x_i - c_j)$

2. ค่าความเป็นสมาชิกของกลุ่มที่ j ของตัวอย่างที่ i คือ $r_{ij} \in \{0,1\}$ ทั้งนี้หาก x_i ถูกกำหนดให้อยู่ในกลุ่มที่ j จะมีค่าเป็น 1 และมีค่าเป็น 0 สำหรับกลุ่มอื่น ๆ ที่ไม่ใช่กลุ่มที่ j กล่าวคือ

$$r_{ij} = \begin{cases} 1, & j = \arg_{\mathrm{m}}\min\|x_i - c_m\|^2 \\ 0, & j \neq \arg_{\mathrm{m}}\min\|x_i - c_m\|^2 \end{cases}$$

ซึ่ง $m = 1,2,\dots,k$ ซึ่งหมายความว่าผลรวมของค่า r_{ij} จะมีค่าเป็น 1 หรือ $\sum_{j=1}^{k} r_{ij} = 1$

การหาค่า c_j ที่เหมาะสมที่สุดสามารถหาได้โดยกำหนดให้อนุพันธ์ย่อยของ $J(r,c)$ เทียบกับ c_j มีค่าเป็นศูนย์ดังนี้

$$\frac{\partial J(r,c)}{\partial c_j} = 2\sum_{i=1}^{n} r_{ij}\left(x_i - c_j\right) = 0$$

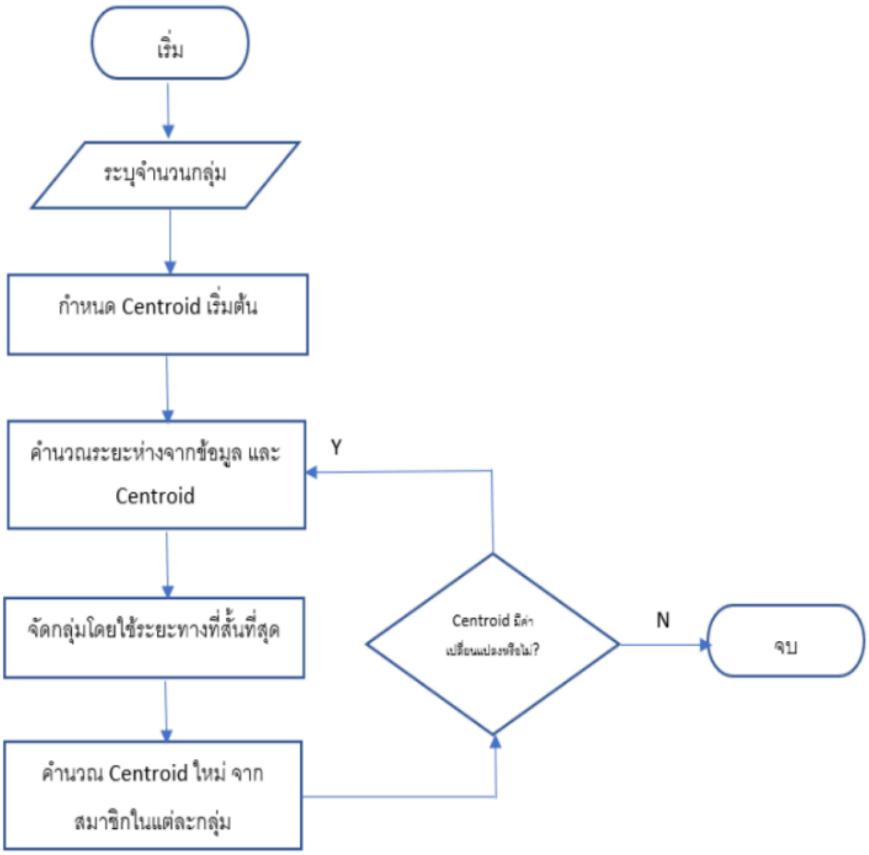
$$\sum_{i=1}^{n} r_{ij}c_j = \sum_{i=1}^{n} r_{ij}x_i$$

$$c_j \sum_{i=1}^{n} r_{ij} = \sum_{i=1}^{n} r_{ij}x_i$$

$$c_j = \frac{\sum_{i=1}^{n} r_{ij}x_i}{\sum_{i=1}^{n} r_{ij}}$$

ซึ่งจะเห็นว่า ตัวหาร หรือ $\sum_{i=1}^{n} r_{ij}$ ก็คือจำนวนตัวอย่างทั้งหมดที่ถูกกำหนดให้อยู่ในกลุ่มที่ j และ c_j ก็คือค่าเฉลี่ย (Mean) ของตัวอย่างทั้งหมดที่ถูกกำหนดให้อยู่ในกลุ่มที่ j นั่นเอง

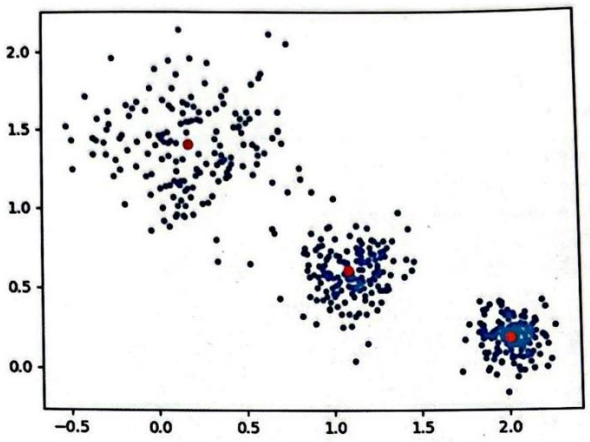
จากผลเฉลยที่ได้จะพบว่า เรายังไม่สามารถหาค่า c_j หรือ Centroid ที่ต้องการได้ เพราะเราไม่ทราบค่า r_{ij} ในขั้นตอนวิธี K-Means แก้ปัญหานี้โดยการกำหนดค่าเริ่มต้นของ Centroid ขึ้นด้วยวิธีการกำหนดค่าเริ่มต้นแบบต่าง ๆ เช่น การกำหนดค่าคงที่ การสุ่มค่า การสุ่มตัวอย่างจากเซตข้อมูล เป็นต้น จากนั้นจะทำการคำนวณระยะห่าง (Distance) จากข้อมูลตัวอย่างทั้งหมดกับ Centroid ของแต่ละกลุ่ม แล้วทำการกำหนดข้อมูลกลุ่มข้อมูลตัวอย่างแต่ละตัวอย่างให้อยู่ในกลุ่มของ Centroid ที่มีระยะห่างที่น้อยที่สุด จากนั้นทำการคำนวณค่า Centroid ของแต่ละกลุ่มใหม่ให้เป็นค่าเฉลี่ยของข้อมูลที่ถูกจัดให้อยู่ในกลุ่มนั้น ๆ จากขั้นตอนก่อนหน้า แล้วทำการคำนวณระยะห่างจากข้อมูลตัวอย่างทั้งหมดกับ Centroid ที่คำนวณขึ้นใหม่ของแต่ละกลุ่มอีกครั้ง แล้วทำการจัดกลุ่มข้อมูลตัวอย่างแต่ละตัวอย่างให้อยู่ในกลุ่มของ Centroid ใหม่ที่มีระยะห่างที่น้อยที่สุดอีกครั้งหนึ่ง จากนั้นจะทำในกระบวนการลักษณะนี้ซ้ำ ๆ จนกว่า Centroid จะไม่มีการเปลี่ยนแปลงหรือมีการเปลี่ยนแปลงน้อย ๆ ดังแสดงในผังงานในรูปที่ 1



รูปที่ 1 ผังงานแสดงขั้นตอนวิธี K-Means

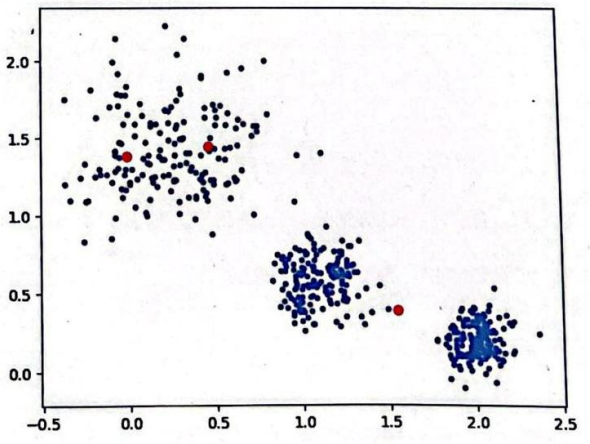
ข้อสังเกต การกำหนดค่าเริ่มต้น (Initailization) ของขั้นตอนวิธี K-Means นอกจากจะทำการกำหนดค่าเริ่มต้นของค่า Centroid หรือ c_j ก่อน อีกวิธีหนึ่งยังสามารถกำหนดค่าเริ่มต้นของ r_{ij} ก่อนแล้วจึงใช้ค่า r_{ij} ที่กำหนดขึ้นนั้นมาคำนวณค่า c_j ต่อไปแทนก็ได้

ในฟังก์ชันนี้ใช้การกำหนดค่าเริ่มต้นของ Centroid โดยทำการสุ่มเลือกจากเซตตัวอย่างจำนวนเท่ากับจำนวนกลุ่ม (k) ที่ต้องการ ส่วนการคำนวณระยะห่างใช้ระยะห่างแบบยูคลิดกำลังสอง เมื่อทำการประมวลผลจะแสดงภาพการเปลี่ยนแปลงของ Centroid ทั้งหมดตั้งแต่เริ่มต้นจนลู่เข้าดังรูปที่ 2



รูปที่ 2 ตัวอย่างการแบ่งกลุ่มข้อมูล 2 มิติด้วยขั้นตอนวิธี K-Means

จากผลลัพธ์ที่ได้จะพบว่า Centroid สามารถลู่เข้าได้ถูกต้อง แต่อย่างไรก็ดี ไม่มีขั้นตอนวิธีจัดกลุ่มใดในปัจจุบันที่รับรองได้ว่า Centroid จะลู่เข้าค่าที่ถูกต้องเสมอ เช่น บางกรณีอาจเกิดการลู่เข้าผิดตำแหน่งดังรูปที่ 3 ทั้งนี้ส่วนใหญ่การกำหนดค่าเริ่มต้นของ Centroid จะเป็นสาเหตุหลักของปัญหานี้



รูปที่ 3 ตัวอย่างการแบ่งกลุ่มข้อมูล 2 มิติด้วยขั้นตอนวิธี K-Means ที่ผิดพลาด

2.8 Standard Scaling

Machine learning algorithm หลายตัวจะทำงานได้ดีเมื่อข้อมูล Input อยู่ใน Scale มาตรฐาน นั่นคือมีค่าเฉลี่ย Mean เท่ากับ 0 และ Variance เท่ากับ 1 ดังนั้นหากเราใช้ Algorithm เหล่านี้ หรือลองเรียนรู้แบบจำลองแล้วได้ค่าความแม่นยำต่ำ หรือใช้เวลานานมากในการเรียนรู้ ให้ลองเปลี่ยน Scale ของตัวแปรดู

Standard scaling มีหลายสูตร แต่สูตรที่ใช้งานได้ดีและเป็นที่นิยม คือสูตร

$$\text{ข้อมูล } scaling \text{ ที่ } i = \frac{\text{ข้อมูลที่ } i - \text{ค่าเฉลี่ย}}{\text{ส่วนเบี่ยงเบนมาตรฐาน}}, i = 1,2,\dots,N$$

ที่ซึ่ง N คือจำนวนข้อมูลทั้งหมด

ค่าเฉลี่ย (Mean) ของข้อมูลทั้งหมด

$$\text{ค่าเฉลี่ย} = \frac{\text{ผลรวมของข้อมูลทั้งหมด}}{\text{จำนวนข้อมูลทั้งหมด}}$$

ส่วนเบี่ยงเบนมาตรฐาน (Standard deviation) ของข้อมูลทั้งหมด

$$\text{ส่วนเบี่ยงเบนมาตรฐาน} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{ข้อมูลที่ } i - \text{ค่าเฉลี่ย})^2}$$

ตัวอย่าง กำหนดให้

ข้อมูลที่ 1	4
ข้อมูลที่ 2	7
ข้อมูลที่ 3	2
ข้อมูลที่ 4	3

$$\text{ค่าเฉลี่ย} = \frac{\text{ผลรวมของข้อมูลทั้งหมด}}{\text{จำนวนข้อมูลทั้งหมด}}$$

$$\text{ค่าเฉลี่ย} = \frac{4+7+2+3}{4} = 4$$

$$\text{ส่วนเบี่ยงเบนมาตรฐาน} = \sqrt{\frac{1}{4}[(4-4)^2+(7-4)^2+(2-4)^2+(3-4)^2]} = \sqrt{\frac{7}{2}} \approx 1.87$$

$$\text{ข้อมูลที่ 1 scaling} = \frac{4-4}{1.87} = 0$$

$$\text{ข้อมูลที่ 2 scaling} \approx \frac{7-4}{1.87} \approx 1.6$$

$$\text{ข้อมูลที่ 3 scaling} \approx \frac{2-4}{1.87} \approx -1.07$$

$$\text{ข้อมูลที่ 4 scaling} \approx \frac{3-4}{1.87} \approx -0.53$$

จะเห็นได้ว่าค่าเฉลี่ยและความแปรปรวนของข้อมูลที่ scaling ทั้งหมดมีค่าเป็น 0 และ 1 ตามลำดับ

2.9 การแบ่งข้อมูล Training และ Test Set

- **Data set** หรือ **Dataset** หมายถึงข้อมูลที่ได้รวบรวมไว้ เพื่อนำมาสอน (Train) ให้กับคอมพิวเตอร์เพื่อสร้างเป็น Model หรือใช้ทดสอบความถูกต้องแม่นยำของ Model คำว่า data set บางทีเรียกว่า ตัวอย่าง/Samples/Instances/observations

- **Training Set/Training Data/Learning data:** ชุดข้อมูลที่นำไปทำการสอนให้กับคอมพิวเตอร์ โดยปกติ จะแบ่ง Data set ออกเป็น 2 ส่วนคือ Training set สำหรับการ Train และ Test set สำหรับทดสอบ

- **Test set:** ชุดข้อมูลที่แบ่งมาจาก Data set เพื่อนำมาทดสอบความแม่นยำ ความถูกต้องของ Model ที่ Train เรียบร้อยแล้ว

สมมติว่ามีข้อมูลอยู่ 10,000 ตัวอย่าง แล้วเอาทั้ง 10,000 ป้อนให้โมเดล Machine Learning ใช้สำหรับ Train ทั้งหมด แล้วเราจะไม่สามารถนำข้อมูลอื่นมาทดสอบการทำงานของโมเดล ทำให้เราไม่ทราบว่าโมเดลทำงานได้แม่นยำมากน้อยเพียงใด ดังนั้นเราจะแก้ปัญหานี้อย่างไร

การที่โมเดลสามารถทำงานได้แม่นยำกับข้อมูลที่ไม่เคยมีมาก่อน เรียกว่า **การทำให้อยู่ในรูปทั่วไป Generalization** เป็นคอนเซ็ปท์สำคัญ ของการพัฒนาระบบ Machine Learning เพราะถ้าเรามีระบบที่ทำงานได้แม่นยำเฉพาะกับข้อมูลได้รับการรวบรวม เปรียบเหมือนกับนักเรียนที่จำข้อสอบ เข้าไปสอบ ทำถูกเฉพาะโจทย์ที่เหมือนเดิมเท่านั้น ไม่สามารถพลิกแพลงกับโจทย์ที่แตกต่างกันได้แม้แต่เล็กน้อยก็ตาม เมื่อเอามาใช้งานจริง เจอข้อมูลจริง ๆ โมเดลก็จะมีประสิทธิภาพความแม่นยำต่ำจนรับไม่ได้ เรียกว่า **โอเวอร์ฟิต Overfit**

เราจึงทำการแบ่งข้อมูล Split ออกเป็น 2 ส่วน คือ Training Set และ Test Set

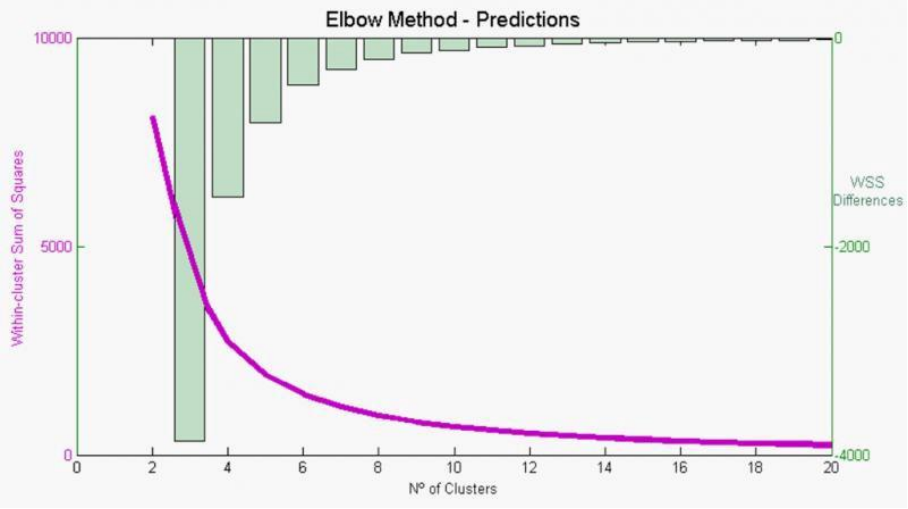
1. **ข้อมูลสำหรับการเรียนรู้ (Training Set)** ใช้สำหรับป้อนให้โมเดลใช้เทรน
2. **ข้อมูลสำหรับการทดสอบ (Test Set)** ใช้สำหรับทดสอบหาความแม่นยำ หลังจากเทรนเสร็จว่าโมเดลจะทำงานได้ดีแค่ไหนกับข้อมูลที่ไม่เคยเห็นมาก่อน

เช่น 9,000 เป็น Training Set และ 1,000 เป็น Test Set

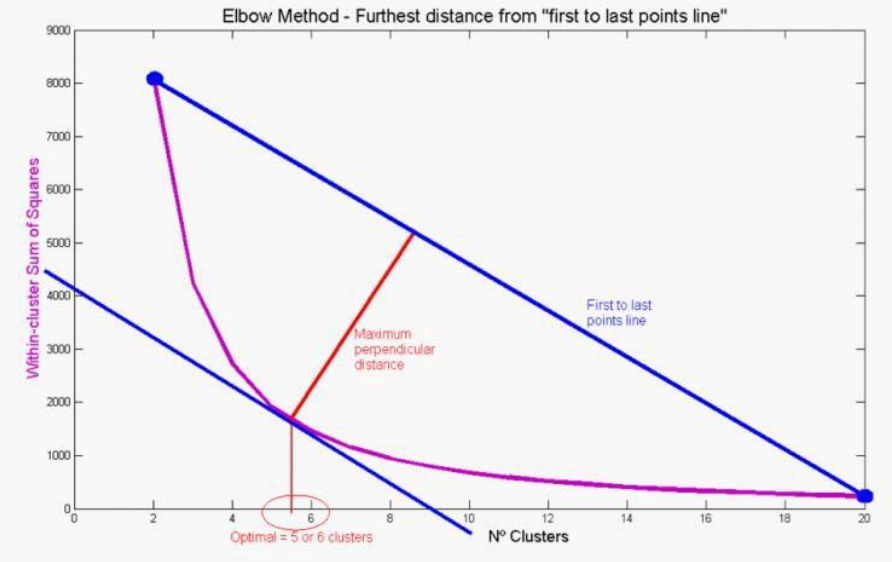
2.10 การหาจำนวน k ที่เหมาะสมที่สุดด้วยวิธี Elbow Method [7]

การเลือกค่า k ที่เหมาะสมหรือการหาค่า Optimal cluster number มีหลายวิธี Elbow method เป็นวิธีหนึ่งซึ่งใช้การวัดข้อผิดพลาด (Error measurement) ของผลรวมของระยะห่างระหว่าง Object กับ Centroid

เมื่อความผิดพลาดลดน้อยลง เส้นโค้งที่มีความชันจะเริ่มโค้งและราบเรียบ (Smooth) จนเกิดเป็นมุมลักษณะเหมือน Elbow ณ จุดนี้เป็นจุดที่ให้ค่าจำนวนกลุ่ม Cluster ที่ดีที่สุด



วิธีการคำนวณหาจุดของเส้นโค้งที่มีจำนวน Cluster ที่เหมาะสมที่สุดนั้น ให้ลากเส้นตรงจากจุดเริ่มไปยังปลายเส้นโค้ง จากนั้นหาระยะจากเส้นตรงตั้งฉากกับเส้นโค้งที่มีระยะห่างมากที่สุดก็จะได้ Optimal cluster number จากตัวอย่างจะเป็นว่าค่าใกล้กับ k = 6



บทที่ 3

วิธีดำเนินโครงงาน

วิธีดำเนินงานค้นคว้าอิสระนี้เริ่มต้นจากการเก็บรวบรวมข้อมูลบุคคลกลุ่มตัวอย่าง อายุ 40 ขึ้นไป โดยวัดค่าข้อมูลจากเครื่องชั่งอัจฉริยะ InBody Dial Body Composition Analyzer เป็นหลัก ประกอบกับใช้ข้อมูลแบบฟอร์มสำรวจโรคของบุคคลกลุ่มตัวอย่างเพื่อวิเคราะห์ จัดกลุ่ม และสังเกตแนวโน้มของโรคไม่ติดต่อที่พบบ่อยที่ได้จากการจัดกลุ่มแบบ Fuzzy C-Means ผ่านโปรแกรมภาษา Python Editor VS Code

3.1 การจัดกลุ่มด้วยวิธี Fuzzy C-Means

การจัดกลุ่มด้วยขั้นตอนวิธี K-Means ในบทที่ 2 หรือที่สามารถเรียกได้อีกชื่อว่า “Hard C-Means” ซึ่งหมายความว่า แต่ละข้อมูลจะถูกจัดให้อยู่ในกลุ่มใดกลุ่มหนึ่งเท่านั้น ทำให้เกิดปัญหาขึ้นกับตัวอย่างที่อยู่บริเวณขอบของแต่ละกลุ่ม ทั้งนี้เนื่องจากตัวอย่างที่ขอบหรือบริเวณใกล้เคียงนั้น ไม่สามารถระบุอย่างชัดเจนได้ว่าเป็นข้อมูลที่อยู่ในกลุ่มใด การแก้ปัญหานี้ทำได้โดยการไม่ระบุให้ตัวอย่างเป็นสมาชิกของกลุ่มใดกลุ่มหนึ่งเพียงกลุ่มเดียว แต่ใช้หลักการตรรกะคลุมเครือ (Fuzzy Logic) มาช่วยในการจัดกลุ่มแทน ซึ่งขั้นตอนวิธีหนึ่งที่นิยมอย่างมากเรียกว่า “ขั้นตอนวิธี Fuzzy C-Means” ขั้นตอนวิธีนี้มีความคล้ายคลึงกับขั้นตอนวิธี K-Means แต่แตกต่างกันที่ข้อมูลแต่ละข้อมูลจะไม่ถูกตัดสินให้อยู่ในกลุ่มใดกลุ่มหนึ่ง แต่จะกำหนดฟังก์ชันความเป็นสมาชิก (Membership Function) ขึ้น เพื่อระบุความเป็นสมาชิกของแต่ละกลุ่ม ดังนั้น ฟังก์ชันวัตถุประสงค์สำหรับการจัดกลุ่มสามารถกำหนดได้ดังนี้

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{k} \mu_{ij}^{p} \left\| x_i - c_j \right\|^2$$

โดย μ_{ik} หมายถึง ค่าความเป็นสมาชิกของตัวอย่าง x_i ในกลุ่มที่ j และ c_j คือ Centroid ของกลุ่มที่ j โดย k เป็นจำนวนกลุ่มทั้งหมด n เป็นจำนวนตัวอย่างทั้งหมด และ p คือ ค่าจำนวนจริงที่มากกว่าหนึ่ง ซึ่งโดยทั่วไปนิยมกำหนดให้มีค่าเป็น 2 ทั้งนี้ผลรวมของค่าความเป็นสมาชิกของทุกกลุ่มสำหรับตัวอย่างหนึ่ง ๆ จะต้องมีค่าเป็น 1 เพื่อให้สามารถเปรียบเทียบระหว่างตัวอย่างได้ดังนี้

$$\sum_{j=1}^{k} \mu_{ij} = 1, \forall i = 1,2,\ldots,n$$

ซึ่งทำให้ค่าความเป็นสมาชิกแต่ละค่าจะมีอยู่ในช่วง $(0,1)$ หรือสามารถคิดเป็นร้อยละหรือความน่าจะเป็นได้

ข้อสังเกต โดยทั่วไป Centroid ที่ดีจะมีระยะห่างจากตัวอย่างภายในกลุ่มน้อย แต่จากฟังก์ชันวัตถุประสงค์จะพบว่าเป็นการหาระยะห่างรวมจาก Centroid ของกลุ่ม ๆ หนึ่งไปยังตัวอย่างทุกตัวอย่างทั้งในและนอกกลุ่มเดียวกันกับ Centroid ทั้งนี้เพราะค่าความเป็นสมาชิก μ จะทำหน้าที่คล้ายค่าน้ำหนักของตัวอย่างนั้น ๆ ซึ่งจะทำให้ค่าระยะห่างเฉลี่ยที่คำนวณได้ของแต่ละ Centroid จะได้รับผลจากตัวอย่างที่อยู่ในกลุ่มมากกว่าผลจากตัวอย่างที่อยู่นอกกลุ่มตามค่าความเป็นสมาชิกดังกล่าว ซึ่งเราสามารถกำหนดผลของค่าสมาชิกได้จากตัวแปร p ซึ่งถ้ากำหนดค่า p มากจะทำให้มีผลน้อย และหากกำหนดค่า p น้อยจะทำให้มีผลมาก

จากฟังก์ชันวัตถุประสงค์ของการจัดกลุ่มด้วยขั้นตอนวิธี Fuzzy C-Means ซึ่งหมายถึงผลรวมของระยะห่างระหว่าง Centroid และตัวอย่างทุกคู่ ซึ่งการจัดกลุ่มที่ดีนั้นจะเกิดขึ้นเมื่อระยะห่างรวมนี้ควรมีค่าน้อยที่สุด ทั้งนี้เพื่อลดความซับซ้อนในการแก้ปัญหานี้ เราสามารถสร้างปัญหาย่อยโดยพิจารณาตัวอย่างเพียงหนึ่งตัวอย่าง เนื่องจากระยะห่างใด ๆ ย่อมมีค่ามากกว่าหรือเท่ากับศูนย์ ดังนั้น ผลรวมของระยะห่างทั้งหมดย่อมแปรผันตรงกับระยะห่างแต่ละระยะห่างด้วย ดังสามารถเป็นปัญหาย่อยเพื่อหาค่าความเป็นสมาชิกที่เหมาะสมที่สุดของตัวอย่างแต่ละตัวอย่างได้ดังนี้

Minimize : $\sum_{j=1}^{k} \mu_{ij}^{p} \left\| x_i - c_j \right\|^2$ (ระยะห่างจาก Centroid น้อยที่สุด)

Subject to : $\sum_{j=1}^{k} \mu_{ij} = 1$ (ค่าความเป็นสมาชิกรวมทุกกลุ่มของตัวอย่างหนึ่งมีค่าเป็น 1)

การแก้ปัญหานี้ทำได้โดยวิธีการตัวคูณลากรองจ์ (Lagrange Multiplier) โดยสามารถสร้างฟังก์ชันลากรองจ์จากปัญหานี้ได้ดังนี้

$$L(\mu, \lambda) = \sum_{j=1}^{k} \mu_{ij}^{p} \left\| x_i - c_j \right\|^2 - \lambda_i \left(\sum_{j=1}^{k} \mu_{ij} - 1 \right)$$

โดย λ คือ ตัวคูณลากรองจ์ ซึ่งสามารถหาค่าความเป็นสมาชิกที่เหมาะสมได้จากค่าเกรเดียนต์ของฟังก์ชันที่มีค่าเป็น 0 หรือ

$$\frac{\partial L(\mu, \lambda)}{\partial \mu} = 0$$

โดยสามารถหาอนุพันธ์ย่อยเทียบกับ μ_{ij} แต่ละค่าได้ดังนี้

$$p\left\|x_i - c_j\right\|^2 \mu_{ij}^{p-1} - \lambda_i = 0$$

$$\mu_{ij}^{p-1} = \frac{\lambda_i}{p\left\|x_i - c_j\right\|^2}$$

$$\mu_{ij} = \left(\frac{\lambda_i}{p\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}}$$

หรือจะได้ค่าความเป็นสมาชิกที่เหมาะสมที่สุดเป็น

$$\mu_{ij} = \left(\frac{\lambda}{p\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}}$$

พจน์ที่มีเทอมของ λ สามารถหาได้โดยการแทนค่าความเป็นสมาชิกที่ได้ในเงื่อนไข จะได้ว่า

$$\sum_{j=1}^{k}\left(\frac{\lambda}{p\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}} = 1$$

$$\left(\frac{\lambda}{p}\right)^{\frac{1}{p-1}} = \frac{1}{\sum_{j=1}^{k}\left(\frac{1}{\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}}}$$

เมื่อแทนเทอม $\left(\frac{\lambda}{p}\right)^{\frac{1}{p-1}}$ จะได้ค่าความเป็นสมาชิกดังนี้

$$\mu_{ij} = \frac{\left(\frac{1}{\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}}}{\sum_{j=1}^{k}\left(\frac{1}{\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}}}$$

เพื่อลดความซ้ำซ้อนในการคำนวณเราสามารถกำหนดให้

$$\mu'_{ij} = \left(\frac{1}{\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}}$$

จากนั้นเมื่อคำนวณค่า μ'_{ij} ได้ครบทุกค่าจึงนำค่าผลรวมมาทำการ Normalization เพื่อหาค่า μ_{ij} ได้ดังนี้

$$\mu_{ij} = \frac{\mu'_{ij}}{\sum_{j=1}^{k} \mu'_{ij}}$$

และหา c_j จาก

$$S(c_j) = \sum_{i=1}^{n} \mu_{ij}^{p} \left\| x_i - c_j \right\|^2$$

หา c_j ได้จากค่าเกรเดียนต์ของฟังก์ชันที่มีค่าเป็น 0 หรือ

$$\frac{\partial S}{\partial c_j} = 0$$

โดยสามารถหาอนุพันธ์ย่อยเทียบกับ c_j ได้ดังนี้

$$\frac{\partial S}{\partial c_j} = -2 \sum_{i=1}^{n} \mu_{ij}^{p} \left\| x_i - c_j \right\| = 0$$

$$\sum_{i=1}^{n} \mu_{ij}^{p} \left\| x_i - c_j \right\| = 0$$

$$\sum_{i=1}^{n} \mu_{ij}^{p} x_i - \sum_{i=1}^{n} \mu_{ij}^{p} c_j = 0$$

$$\sum_{i=1}^{n} \mu_{ij}^{p} c_j = \sum_{i=1}^{n} \mu_{ij}^{p} x_i$$

$$c_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{p} x_i}{\sum_{i=1}^{n} \mu_{ij}^{p}}$$

<u>วิธีแบ่งกลุ่มแบบ Fuzzy C-Means</u>

Step 1 สุ่ม μ_{ij} โดยอยู่ในเงื่อนไข

$$\sum_{j=1}^{k} \mu_{ij} = 1$$

Step 2 หา Centroids จาก

$$c_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{p} \, x_i}{\sum_{i=1}^{n} \mu_{ij}^{p}}$$

Step 3 หาค่าความเป็นสมาชิกจาก

$$\mu_{ij} = \frac{\left(\frac{1}{\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}}}{\sum_{j=1}^{k}\left(\frac{1}{\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}}}$$

Step 4 ทำ Step 2 และ Step 3 ซ้ำไปเรื่อย ๆ จน Centroids ไม่เปลี่ยนแปลง

<u>ตัวอย่าง</u> ข้อมูล $(1,2),(2,3),(9,4),(10,1),\ \ k=2,p=2$

x_1	1	2
x_2	2	3
x_3	9	4
x_4	10	1

<u>Step 1</u> สุ่ม μ_{ij}

μ_{ij}	c_1	c_2
x_1	0.40	0.60
x_2	0.88	0.12
x_3	0.41	0.59
x_4	0.27	0.73

<u>Step 2</u> หา Centroids จาก $c_j = \frac{\sum_{i=1}^{n}\mu_{ij}^{p}x_i}{\sum_{i=1}^{n}\mu_{ij}^{p}}$

$$\sum_{i=1}^{n}\mu_{i1}^2 = (0.40)^2 + (0.88)^2 + (0.41)^2 + (0.27)^2 = 1.18$$

$$\sum_{i=1}^{n}\mu_{i2}^2 = (0.60)^2 + (0.12)^2 + (0.59)^2 + (0.73)^2 = 1.25$$

หา Centroids ที่ 1 : $c_1 = [c_{11}, c_{12}]$

$$c_{11} = \frac{(0.40)^2 \times 1 + (0.88)^2 \times 2 + (0.41)^2 \times 9 + (0.27)^2 \times 10}{1.18} = 3.38$$

$$c_{12} = \frac{(0.60)^2 \times 2 + (0.12)^2 \times 3 + (0.59)^2 \times 4 + (0.73)^2 \times 1}{1.18} = 2.88$$

ดังนั้น $c_1 = [3.38,2.88]$

หา Centroids ที่ 2 : $c_2 = [c_{21}, c_{22}]$

$$c_{21} = 7.02$$

$$c_{22} = 2.14$$

ดังนั้น $c_2 = [7.02,2.14]$

Step 3 หาค่าความเป็นสมาชิก

x_1	1	2
x_2	2	3
x_3	9	4
x_4	10	1

c_1	3.38	2.88
c_2	7.02	2.14

$\|\|x_i - c_j\|\|^2$	c_1	c_2
x_1	6.44	36.26
x_2	1.92	25.94
x_3	32.84	7.38
x_4	47.36	10.18

จาก

$$\mu_{ij} = \frac{\left(\frac{1}{\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}}}{\sum_{j=1}^{k}\left(\frac{1}{\left\|x_i - c_j\right\|^2}\right)^{\frac{1}{p-1}}}$$

จะได้

$$\mu_{11} = \frac{\left(\frac{1}{\|x_1 - c_1\|^2}\right)}{\left(\frac{1}{\|x_1 - c_1\|^2}\right) + \left(\frac{1}{\|x_1 - c_2\|^2}\right)} = \frac{\frac{1}{6.44}}{\frac{1}{6.44} + \frac{1}{36.26}} = \frac{0.16}{0.18} = 0.85$$

$$\mu_{12} = \frac{\left(\frac{1}{\|x_1 - c_2\|^2}\right)}{\left(\frac{1}{\|x_1 - c_1\|^2}\right) + \left(\frac{1}{\|x_1 - c_2\|^2}\right)} = \frac{\frac{1}{36.26}}{\frac{1}{6.44} + \frac{1}{36.26}} = \frac{0.03}{0.18} = 0.15$$

ดังนั้น

μ_{ij}	c_1	c_2
x_1	0.85	0.15
x_2	0.93	0.07
x_3	0.18	0.82
x_4	0.18	0.82

Step 4 ทำ Step 2 และ Step 3 ซ้ำไปเรื่อย ๆ จน Centroids ไม่เปลี่ยนแปลง

ได้ว่า

	c_1	c_2	
x_1	0.88	0.12	c_1
x_2	0.94	0.06	c_1
x_3	0.17	0.83	c_2
x_4	0.18	0.82	c_2

3.2 การเก็บรวบรวมข้อมูล

กลุ่มตัวอย่างที่ใช้ในการศึกษาคือ บุคคลเพศชายและเพศหญิงอายุ 40 ปีขึ้นไป จำนวน 50 คน แบ่งเป็น 2 กลุ่มคือ จำนวน 40 คน คิดเป็น 80% เป็นข้อมูลสำหรับการเรียนรู้ของแบบจำลอง และจำนวน 10 คน คิดเป็น 20% เป็นข้อมูลสำหรับการทดสอบแบบจำลอง

ข้อมูลที่นำมาวิเคราะห์ในการศึกษา แบ่งเป็น 2 กลุ่มคือ

1. ค่าที่วัดได้จากเครื่องชั่ง InBody Dial Body Composition Analyzer คือ

1.1 Weight (น้ำหนัก) หน่วย kg.

1.2 Body Fat (ไขมันในร่างกาย) หน่วย %

ไขมันในร่างกาย (Body Fat Percent หรือ Body Fat Percentage) คือ สัดส่วนของไขมันในร่างกายของเรา (ทั้งไขมันจำเป็นและไขมันส่วนเกิน) ซึ่งจะคิดเป็นร้อยละหรือเปอร์เซ็นต์เมื่อเทียบกับน้ำหนักร่างกาย เช่น หากเราหนัก 50 กิโลกรัม และมีปริมาณไขมันในร่างกายหนัก 10 กิโลกรัม ก็แสดงว่ามีไขมันในร่างกาย หรือ Body Fat Percent 20% นั่นเอง

เกณฑ์ Body Fat Percent ที่เหมาะสม คือ ชาย 13-20% และหญิง 23-30% ซึ่งอาจแตกต่างกันออกไปแต่จะมีความใกล้เคียงกัน และโดยทั่วไปปริมาณไขมันแต่ละระดับสามารถพิจารณาได้ดังนี้

เปอร์เซ็นต์ไขมัน	ลักษณะ
ชาย 30% ขึ้นไป หญิง 40% ขึ้นไป	ปริมาณไขมันมากในระดับวิกฤต ทำให้รูปร่างอ้วนกลม มองเห็นเซลลูไลท์บนผิวหนังได้ชัดเจน และมีไขมันส่วนเกินกระจายอยู่ทุกส่วนของร่างกาย รูปร่างมองดูเป็นชั้นๆ ชัดเจน
ชาย 21-30% หญิง 31-40%	มีปริมาณไขมันส่วนเกินในร่างกาย มีรูปร่างท้วม และมีชั้นไขมันหนาหุ้มกล้ามเนื้ออยู่
ชาย 13-20% หญิง 23-30%	เป็นปริมาณไขมันตามมาตรฐานทั่วไป สุขภาพอยู่ในเกณฑ์ดี มีร่างกายสมส่วน เป็นสัดส่วนชัดเจน แม้ว่ายังไม่เห็นกล้ามเนื้อมากนัก

ชาย 9-12 % หญิง 19-22%	มีปริมาณไขมันน้อย มีกล้ามเนื้อชัดเจนมากขึ้น รูปร่างกระชับสมส่วน ในผู้ชายที่ออกกำลังกายจะมีกล้ามเนื้อหน้าท้อง (Six Packs) ชัดเจนขึ้นเรื่อยๆ
ชาย 5-8% หญิง 15-18%	ปริมาณไขมันน้อยมาก เห็นกล้ามเนื้อชัดเจน รูปร่างเพรียวบาง และเป็นปริมาณไขมันที่ผู้หญิงสามารถออกกำลังกายเพิ่มกล้ามเนื้อหน้าท้อง (Six Packs) ได้
ชาย 5-8% หญิง 15-18%	ปริมาณไขมันน้อยมาก เห็นกล้ามเนื้อชัดเจน รูปร่างเพรียวบาง และเป็นปริมาณไขมันที่ผู้หญิงสามารถออกกำลังกายเพิ่มกล้ามเนื้อหน้าท้อง (Six Packs) ได้
ชาย น้อยกว่า 5% หญิง น้อยกว่า 15%	ปริมาณไขมันน้อยจนถึงขั้นวิกฤต รูปร่างผอมบางมากเกินไป อาจทำให้สุขภาพอ่อนแอ และเป็นอันตรายต่อระบบการทำงานของอวัยวะต่างๆ ภายในร่างกายได้

1.3 Muscle (มวลกล้ามเนื้อ) หน่วย %

มวลกล้ามเนื้อ (Muscle) คือ น้ำหนักรวมของกล้ามเนื้อในร่างกายคนเรา โดยไม่ได้นับรวมไขมัน เส้นเอ็น และกระดูก มวลกล้ามเนื้อนี้ไม่ใช่น้ำหนักตัว บางคนน้ำหนักตัวเยอะ แต่กล้ามเนื้อน้อย แล้วไปเยอะที่ไขมันแทน จากผลงานวิจัยพบกว่า เมื่ออายุมากขึ้น ทุกคน ทุกเพศ จะมีมวลกล้ามเนื้อลดลงเรื่อย ๆ

- ค่ามวลกล้ามเนื้อในเพศหญิงคิดเป็นเปอร์เซ็นต์ของมวลกาย

อายุ	ค่ามวลกล้ามเนื้อปกติ
18-40 ปี	24.4-30.2%
41-60 ปี	24.2-30.3%
61-80 ปี	24.0-29.8%

- ค่ามวลกล้ามเนื้อในเพศชายคิดเป็นเปอร์เซ็นต์ของมวลกาย

อายุ	ค่ามวลกล้ามเนื้อปกติ
18-40 ปี	33.4-39.4%
41-60 ปี	33.2-39.2%
61-80 ปี	33.0-38.7%

1.4 Visceral Fat (ไขมันในช่องท้อง)

ไขมันในช่องท้อง (Visceral Fat) คือ การที่ร่างกายรับสารอาหารประเภทไขมัน รวมไปถึงคาร์โบไฮเดรตและน้ำตาลในปริมาณที่มากเกินไป จนทำให้ร่างกายเผาผลาญไม่หมด สุดท้ายก็จะถูกเปลี่ยนสภาพมาเป็นไขมัน โดยไขมันในช่องท้องจะสะสมอยู่ลึกกว่าชั้นผิวหนัง สะสมอยู่รอบอวัยวะภายในร่างกาย เช่น กระเพาะอาหาร ตับ หรือลำไส้เล็ก ซึ่งตับอาจเปลี่ยนไขมันนี้เป็นคอเลสเตอรอล รวมทั้งอาจดูดซึมเข้ากระแสเลือดและสะสมตามผนังหลอดเลือดแดง ทำให้หลอดเลือดตีบ ไขมันช่องท้องก่อให้เกิดปัญหาสุขภาพต่าง ๆ

ระดับไขมันในช่องท้อง	ลักษณะ
1-9	อยู่ในระดับปกติ
10-14	เริ่มมีความเสี่ยง
15 ขึ้นไป	มีความเสี่ยงสูง

2. ข้อมูลทั่วไป ส่วนสูง, เพศ, อายุ และโรคประจำตัว

ซึ่งข้อมูลนี้ได้เก็บรวบรวมจากแบบฟอร์มเก็บข้อมูล ดังรูป 4

แบบฟอร์มเก็บข้อมูลสำหรับงานวิจัยเรื่อง Fuzzy C-Means Clustering Based on Body Composition Scale for Analysis of Common Elderly Illnesses

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

ชื่อ – นามสกุล ..... **อายุ** ..... **เบอร์โทร** .....

เพศ ☐ ชาย ☐ หญิง

Height (cm) ..... Weight (kg) ..... Body Fat (%) .....

Muscle (kg) ..... Visceral Fat .....

โรคประจำตัว

☐ โรคมะเร็ง ☐ โรคอ้วน ☐ โรคหัวใจ ☐ โรคเบาหวาน

☐ โรคความดันโลหิตสูง ☐ โรคตับ ☐ โรคไต ☐ โรคไขมันในเลือดสูง

☐ อื่นๆ .....

ลงชื่อ .....

(.....)

ผู้ให้ข้อมูล

วันที่ ...../...../.....

ลงชื่อ .....

(.....)

ผู้จัดเก็บข้อมูล

วันที่ ...../...../.....

รูปที่ 4 แบบฟอร์มเก็บข้อมูลสำหรับงานวิจัยนี้

3.3 การเตรียมข้อมูล

1. รวบรวมข้อมูลที่วัดได้จากเครื่องชั่งของกลุ่มตัวอย่างลงใน excel โดยข้อมูลมีดังนี้

คอลัมน์ A : เพศ (เพศหญิง = 1, เพศชาย = 2)

คอลัมน์ B : อายุ (อายุ ตั้งแต่ 40 ปีขึ้นไป)

คอลัมน์ C : ส่วนสูง (Height)

คอลัมน์ D : น้ำหนัก (Weight)

คอลัมน์ E : ไขมันในร่างกาย (Body Fat)

คอลัมน์ F : กล้ามเนื้อ (Muscle)

คอลัมน์ G : ไขมันในช่องท้อง (Visceral Fat)

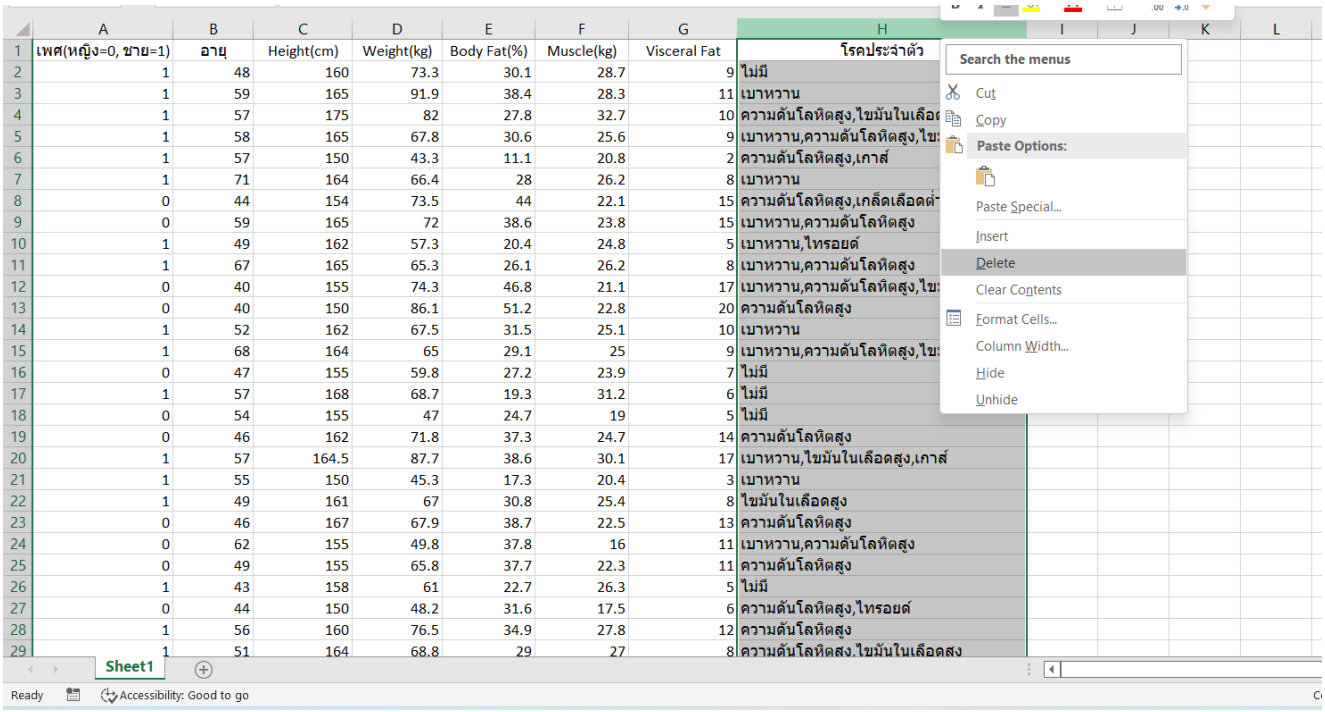
คอลัมน์ H : โรคประจำตัว

	A	B	C	D	E	F	G	H	I
1	เพศ(หญิง=0, ชาย=1)	อายุ	Height(cm)	Weight(kg)	Body Fat(%)	Muscle(kg)	Visceral Fat	โรคประจำตัว	
2	1	48	160	73.3	30.1	28.7	9	ไม่มี	
3	1	59	165	91.9	38.4	28.3	11	เบาหวาน	
4	1	57	175	82	27.8	32.7	10	ความดันโลหิตสูง,ไขมันในเลือดสูง	
5	1	58	165	67.8	30.6	25.6	9	เบาหวาน,ความดันโลหิตสูง,ไขมันในเลือดสูง	
6	1	57	150	43.3	11.1	20.8	2	ความดันโลหิตสูง,เกาส์	
7	1	71	164	66.4	28	26.2	8	เบาหวาน	
8	0	44	154	73.5	44	22.1	15	ความดันโลหิตสูง,เกล็ดเลือดต่ำ	
9	0	59	165	72	38.6	23.8	15	เบาหวาน,ความดันโลหิตสูง	
10	1	49	162	57.3	20.4	24.8	5	เบาหวาน,ไทรอยด์	
11	1	67	165	65.3	26.1	26.2	8	เบาหวาน,ความดันโลหิตสูง	
12	0	40	155	74.3	46.8	21.1	17	เบาหวาน,ความดันโลหิตสูง,ไขมันในเลือดสูง	
13	0	40	150	86.1	51.2	22.8	20	ความดันโลหิตสูง	
14	1	52	162	67.5	31.5	25.1	10	เบาหวาน	
15	1	68	164	65	29.1	25	9	เบาหวาน,ความดันโลหิตสูง,ไขมันในเลือดสูง	
16	0	47	155	59.8	27.2	23.9	7	ไม่มี	
17	1	57	168	68.7	19.3	31.2	6	ไม่มี	
18	0	54	155	47	24.7	19	5	ไม่มี	
19	0	46	162	71.8	37.3	24.7	14	ความดันโลหิตสูง	
20	1	57	164.5	87.7	38.6	30.1	17	เบาหวาน,ไขมันในเลือดสูง,เกาส์	
21	1	55	150	45.3	17.3	20.4	3	เบาหวาน	
22	1	49	161	67	30.8	25.4	8	ไขมันในเลือดสูง	
23	0	46	167	67.9	38.7	22.5	13	ความดันโลหิตสูง	
24	0	62	155	49.8	37.8	16	11	เบาหวาน,ความดันโลหิตสูง	
25	0	49	155	65.8	37.7	22.3	11	ความดันโลหิตสูง	
26	1	43	158	61	22.7	26.3	5	ไม่มี	
27	0	44	150	48.2	31.6	17.5	6	ความดันโลหิตสูง,ไทรอยด์	
28	1	56	160	76.5	34.9	27.8	12	ความดันโลหิตสูง	
29	1	51	164	68.8	29	27	8	ความดันโลหิตสูง,ไขมันในเลือดสูง	

Sheet1

Ready Accessibility: Good to go

2. ลบคอลัมน์ H (โรคประจำตัว) ออก เนื่องจากไม่ได้นำมาใช้ในการคำนวณ Python แต่ยังคงคอลัมน์ A ถึง G ดังเดิม



	A	B	C	D	E	F	G	H
1	เพศ(หญิง=0, ชาย=1)	อายุ	Height(cm)	Weight(kg)	Body Fat(%)	Muscle(kg)	Visceral Fat	โรคประจำตัว
2	1	48	160	73.3	30.1	28.7	9	ไม่มี
3	1	59	165	91.9	38.4	28.3	11	เบาหวาน
4	1	57	175	82	27.8	32.7	10	ความดันโลหิตสูง,ไขมันในเลือด
5	1	58	165	67.8	30.6	25.6	9	เบาหวาน,ความดันโลหิตสูง,ไข
6	1	57	150	43.3	11.1	20.8	2	ความดันโลหิตสูง,เกาส์
7	1	71	164	66.4	28	26.2	8	เบาหวาน
8	0	44	154	73.5	44	22.1	15	ความดันโลหิตสูง,เกล็ดเลือดต่
9	0	59	165	72	38.6	23.8	15	เบาหวาน,ความดันโลหิตสูง
10	1	49	162	57.3	20.4	24.8	5	เบาหวาน,ไทรอยด์
11	1	67	165	65.3	26.1	26.2	8	เบาหวาน,ความดันโลหิตสูง
12	0	40	155	74.3	46.8	21.1	17	เบาหวาน,ความดันโลหิตสูง,ไข
13	0	40	150	86.1	51.2	22.8	20	ความดันโลหิตสูง
14	1	52	162	67.5	31.5	25.1	10	เบาหวาน
15	1	68	164	65	29.1	25	9	เบาหวาน,ความดันโลหิตสูง,ไข
16	0	47	155	59.8	27.2	23.9	7	ไม่มี
17	1	57	168	68.7	19.3	31.2	6	ไม่มี
18	0	54	155	47	24.7	19	5	ไม่มี
19	0	46	162	71.8	37.3	24.7	14	ความดันโลหิตสูง
20	1	57	164.5	87.7	38.6	30.1	17	เบาหวาน,ไขมันในเลือดสูง,เกาส์
21	1	55	150	45.3	17.3	20.4	3	เบาหวาน
22	1	49	161	67	30.8	25.4	8	ไขมันในเลือดสูง
23	0	46	167	67.9	38.7	22.5	13	ความดันโลหิตสูง
24	0	62	155	49.8	37.8	16	11	เบาหวาน,ความดันโลหิตสูง
25	0	49	155	65.8	37.7	22.3	11	ความดันโลหิตสูง
26	1	43	158	61	22.7	26.3	5	ไม่มี
27	0	44	150	48.2	31.6	17.5	6	ความดันโลหิตสูง,ไทรอยด์
28	1	56	160	76.5	34.9	27.8	12	ความดันโลหิตสูง
29	1	51	164	68.8	29	27	8	ความดันโลหิตสูง,ไขมันในเลือดสูง

3. เปลี่ยนชื่อคอลัมน์ A ถึง G เพื่อให้ง่ายในการคำนวณใน Python ดังนี้

คอลัมน์ A : ให้ S แทน เพศ

คอลัมน์ B : ให้ A แทน อายุ

คอลัมน์ C : ให้ H แทน ส่วนสูง (Height)

คอลัมน์ D : ให้ W แทน น้ำหนัก (Weight)

คอลัมน์ E : ให้ BF แทน ไขมันในร่างกาย (Body Fat)

คอลัมน์ F : ให้ M แทน กล้ามเนื้อ (Muscle)

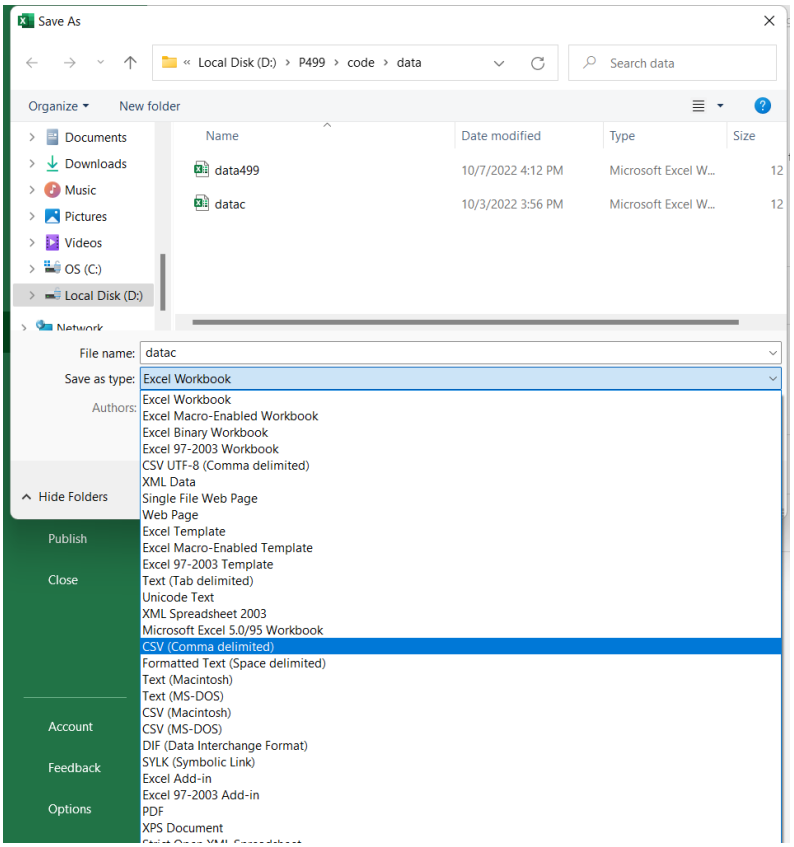
คอลัมน์ G : ให้ VF แทน ไขมันในช่องท้อง (Visceral Fat)

	A	B	C	D	E	F	G
1	S	A	H	W	BF	M	VF
2	1	48	160	73.3	30.1	28.7	9
3	1	59	165	91.9	38.4	28.3	11
4	1	57	175	82	27.8	32.7	10
5	1	58	165	67.8	30.6	25.6	9
6	1	57	150	43.3	11.1	20.8	2
7	1	71	164	66.4	28	26.2	8
8	0	44	154	73.5	44	22.1	15
9	0	59	165	72	38.6	23.8	15
10	1	49	162	57.3	20.4	24.8	5
11	1	67	165	65.3	26.1	26.2	8
12	0	40	155	74.3	46.8	21.1	17
13	0	40	150	86.1	51.2	22.8	20
14	1	52	162	67.5	31.5	25.1	10
15	1	68	164	65	29.1	25	9
16	0	47	155	59.8	27.2	23.9	7
17	1	57	168	68.7	19.3	31.2	6
18	0	54	155	47	24.7	19	5
19	0	46	162	71.8	37.3	24.7	14
20	1	57	164.5	87.7	38.6	30.1	17
21	1	55	150	45.3	17.3	20.4	3
22	1	49	161	67	30.8	25.4	8
23	0	46	167	67.9	38.7	22.5	13
24	0	62	155	49.8	37.8	16	11
25	0	49	155	65.8	37.7	22.3	11
26	1	43	158	61	22.7	26.3	5
27	0	44	150	48.2	31.6	17.5	6
28	1	56	160	76.5	34.9	27.8	12
29	1	51	164	68.8	29	27	8

Sheet1

Ready Accessibility: Good to go

4. Save As ข้อมูล โดยตั้งไฟล์เป็น CSV (Comma delimited) เพื่อให้ข้อมูลใน excel เชื่อมกับ Python โดยอัตโนมัติ



3.4 การสร้างแบบจำลองข้อมูล

1. อ่านไฟล์ข้อมูลเข้าโปรแกรม Python

```
data = pd.read_csv('D:\P499\code\data\datac.csv')
data = pd.DataFrame(data,columns=['S','A','H','W','BF','M','VF'])
```

2. แบ่งข้อมูลออกเป็น 2 ชุด คือ

1) x.train ข้อมูลสำหรับการเรียนรู้ จำนวน 40 ข้อมูล คิดเป็น 80%

2) x.test ข้อมูลสำหรับการทดสอบ จำนวน 10 ข้อมูล คิดเป็น 20%

```
#Spliting Data To X_train and X-test
X_train = data.iloc[:40,:]
X_test = data.iloc[40:,:]
```

3. ปรับข้อมูลในรูปแบบบรรทัดฐาน

```
#Scaling Data
scalarModel = StandardScaler()
X_train = scalarModel.fit_transform(X_train)
X_test = scalarModel.fit_transform(X_test)
```

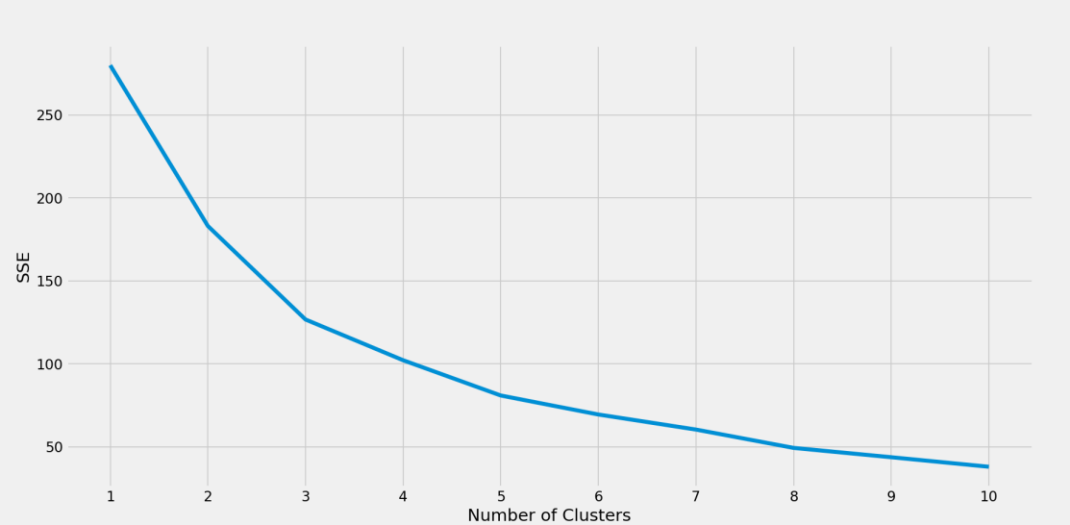
4. หาจำนวนกลุ่มที่เหมาะสมกับชุดข้อมูล x.train

```
#Choosing the Appropriate Number of Clusters
# A list holds the SSE values for each k
sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(X_train)
    sse.append(kmeans.inertia_)

plt.style.use("fivethirtyeight")
plt.plot(range(1, 11), sse)
plt.xticks(range(1, 11))
plt.xlabel("Number of Clusters")
plt.ylabel("SSE")
plt.close()
```

```
kl = KneeLocator(
    range(1, 11), sse, curve="convex", direction="decreasing"
)

print(kl.elbow)
```



5. หาจุดศูนย์กลางของกลุ่มด้วยวิธี Fuzzy C-Means [8]

```
#Apply the algorithm Fuzzy c Means
fcmModel = FCM(n_clusters =kl.elbow)
fcmModel.fit(X_train)
center = fcmModel.centers
print(center)
```

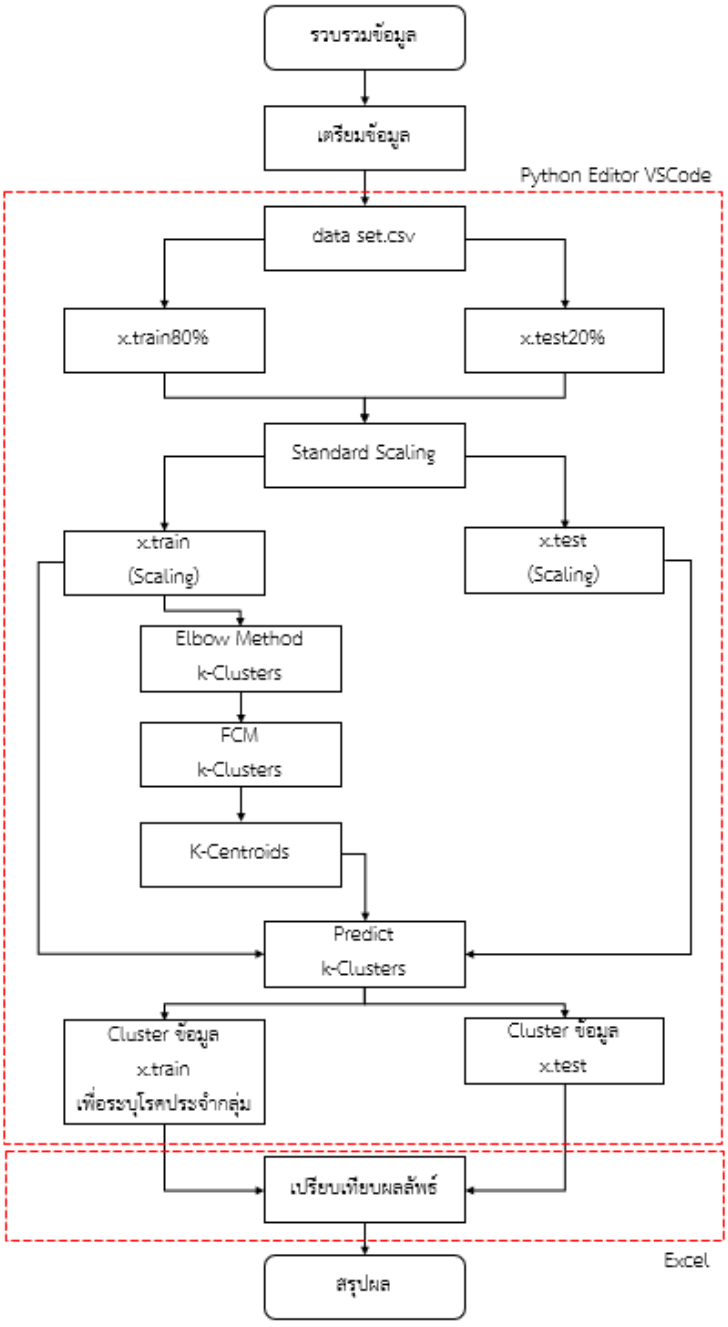
6. ทดสอบแบบจำลองข้อมูลกับข้อมูลชุดทดสอบ

```
#Calculating Prediction
ptrain = fcmModel.predict(X_train)
ptest = fcmModel.predict(X_test)
```

7. แสดงค่าความเป็นสมาชิกใน excel

```
#Calculating Membership Function
u = pd.DataFrame(fcmModel.u,columns=['0','1','2'])
writer = pd.ExcelWriter('solutions.xlsx')
u.to_excel(writer, sheet_name='Sheet1')
writer.save()
```

แผนผังแบบจำลองการแบ่งกลุ่มด้วยวิธี Fuzzy C-Means ผ่านโปรแกรมภาษา Python Editor VSCode ให้เห็นขั้นตอน Preprocessing, Processing และ Postprocessing



รูป 5 แสดงถึงแผนผังแบบจำลองการแบ่งกลุ่มด้วยวิธี Fuzzy C-Means ผ่านโปรแกรมภาษา Python Editor VSCode

3.5 วิเคราะห์โรคไม่ติดต่อที่พบบ่อยเทียบกับผลลัพธ์ที่ได้จากแบบจำลอง Fuzzy C-Means แบบระบุกลุ่มชัดเจน

1. จากการคำนวณผ่านแบบจำลองด้วยโปรแกรมภาษา Python ได้ว่า

1.1 จำนวนกลุ่มที่เหมาะสมที่สุดของข้อมูลสำหรับการเรียนรู้ คือ k = 3

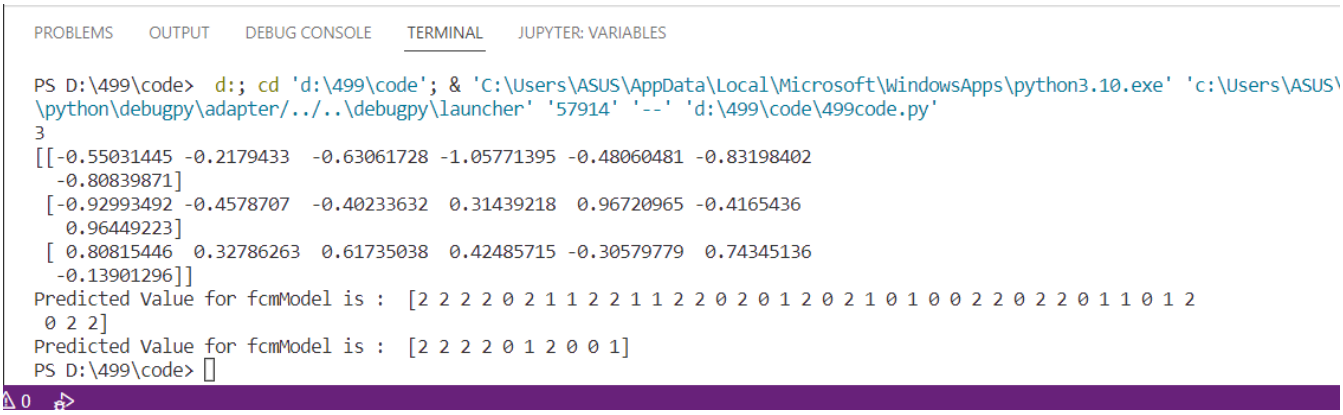
1.2 ค่าจุด Centroid ทั้ง 3 จุด คือ

c_0 =[-0.55031529 -0.21794345 -0.63061757 -1.05771415 -0.48060423 -0.83198459 -0.80839848]

c_1 =[-0.92993465 -0.45787056 -0.40233603 0.3143926 0.96720962 -0.4165432 0.96449244]

c_2 =[0.8081545 0.32786242 0.61735036 0.42485686 -0.30579808 0.74345128 -0.13901327]

โดยที่ c_i คือ จุดศูนย์กลางข้อมูลประจำกลุ่มที่ i , $i = 0,1,2$



2. นำเลขประจำกลุ่มมาใส่ในข้อมูลของเราใน excel

	A	B	C	D	E	F	G	H	I	J	K	L
1	เพศ(หญิง=0, ชาย=1)	อายุ	Height(cm)	Weight(kg)	Body Fat(%)	Muscle(kg)	Visceral Fat	โรคประจำตัว				
2	1	48	160	73.3	30.1	28.7	9	ไม่มี			2	
3	1	59	165	91.9	38.4	28.3	11	เบาหวาน			2	
4	1	57	175	82	27.8	32.7	10	ความดันโลหิตสูง	ไขมันในเลือดสูง		2	
5	1	58	165	67.8	30.6	25.6	9	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	2	
6	1	57	150	43.3	11.1	20.8	2	ความดันโลหิตสูง	เกาส์		0	
7	1	71	164	66.4	28	26.2	8	เบาหวาน			2	
8	0	44	154	73.5	44	22.1	15	ความดันโลหิตสูง			1	
9	0	59	165	72	38.6	23.8	15	เบาหวาน	ความดันโลหิตสูง		1	
10	1	49	162	57.3	20.4	24.8	5	เบาหวาน	ไทรอยด์		2	
11	1	67	165	65.3	26.1	26.2	8	เบาหวาน	ความดันโลหิตสูง		2	
12	0	40	155	74.3	46.8	21.1	17	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	1	
13	0	40	150	86.1	51.2	22.8	20	ความดันโลหิตสูง			1	
14	1	52	162	67.5	31.5	25.1	10	เบาหวาน			2	
15	1	68	164	65	29.1	25	9	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	2	
16	0	47	155	59.8	27.2	23.9	7	ไม่มี			0	
17	1	57	168	68.7	19.3	31.2	6	ไม่มี			2	
18	0	54	155	47	24.7	19	5	ไม่มี			0	
19	0	46	162	71.8	37.3	24.7	14	ความดันโลหิตสูง			1	
20	1	57	164.5	87.7	38.6	30.1	17	เบาหวาน	ไขมันในเลือดสูง	เกาส์	2	
21	1	55	150	45.3	17.3	20.4	3	เบาหวาน			0	
22	1	49	161	67	30.8	25.4	8	ไขมันในเลือดสูง			2	
23	0	46	167	67.9	38.7	22.5	13	ความดันโลหิตสูง			1	
24	0	62	155	49.8	37.8	16	11	เบาหวาน	ความดันโลหิตสูง		0	
25	0	49	155	65.8	37.7	22.3	11	ความดันโลหิตสูง			1	
26	1	43	158	61	22.7	26.3	5	ไม่มี			0	
27	0	44	150	48.2	31.6	17.5	6	ความดันโลหิตสูง	ไทรอยด์		0	
28	1	56	160	76.5	34.9	27.8	12	ความดันโลหิตสูง			2	
29	1	51	164	68.8	29	27	8	ความดันโลหิตสูง	ไขมันในเลือดสูง		2	
30	0	42	150	41.4	26.4	15.9	4	ไม่มี			0	
31	1	42	185	76.5	22.2	33.1	6	ไม่มี			2	
32	1	44	170	67.5	25.8	27.9	7	ความดันโลหิตสูง			2	
33	0	47	158	50.8	29.5	19.2	5	ไม่มี			0	
34	0	47	155	70	36	22	13	ความดันโลหิตสูง	เบาหวาน	ไขมันในเลือด	1	
35	0	48	155	56.4	41	19	11	ความดันโลหิตสูง	เบาหวาน		1	
36	1	43	164	55.5	19.3	24.7	4	ไม่มี			0	
37	0	45	155	55	40.2	17.7	11	ไม่มี			1	
38	1	45	175	75.5	24.8	31.6	9	เบาหวาน	ความดันโลหิตสูง	เกาส์	2	
39	0	47	163	55.7	29.7	21.1	6	ไม่มี			0	
40	1	49	175	71.6	18.9	32.5	5	ไม่มี			2	
41	1	55	163	88	39.5	30	15	ความดันโลหิตสูง	เบาหวาน	ไขมันในเลือด	2	
42	1	53	165	68.9	32	26	10	ความดันโลหิตสูง	ไขมันในเลือดสูง		2	
43	1	44	170	66	28	27	7	ความดันโลหิตสูง			2	
44	0	45	170	78.8	29	32.1	9	ความดันโลหิตสูง	เบาหวาน		2	
45	1	58	165	89.7	38.9	30	16	เบาหวาน	ไขมันในเลือดสูง		2	
46	1	49	159	68	30.6	24.4	9	ไม่มี			0	
47	0	45	164	68.9	39	23.4	14	ความดันโลหิตสูง			1	
48	1	41	173	82.7	29.7	33	9	ไม่มี			2	
49	0	42	157	46	32.6	16.1	7	ไม่มี			0	
50	0	45	161	49.9	33	17.4	7	ไม่มี			0	
51	0	46	164	87	49.5	24.5	20	ไม่มี			1	
52												
53												

3. แยกข้อมูล x.train แต่ละกลุ่มไว้ด้วยกัน

กลุ่ม 0

	A	B	C	D	E	F	G	H	I	J	K	L
1	กลุ่ม 0											
2	1	57	150	43.3	11.1	20.8	2	ความดันโลหิตสูง	เกาส์		0	
3	0	47	155	59.8	27.2	23.9	7	ไม่มี			0	
4	0	54	155	47	24.7	19	5	ไม่มี			0	
5	1	55	150	45.3	17.3	20.4	3	เบาหวาน			0	
6	0	62	155	49.8	37.8	16	11	เบาหวาน	ความดันโลหิตสูง		0	
7	1	43	158	61	22.7	26.3	5	ไม่มี			0	
8	0	44	150	48.2	31.6	17.5	6	ความดันโลหิตสูง	ไทรอยด์		0	
9	0	42	150	41.4	26.4	15.9	4	ไม่มี			0	
10	0	47	158	50.8	29.5	19.2	5	ไม่มี			0	
11	1	43	164	55.5	19.3	24.7	4	ไม่มี			0	
12	0	47	163	55.7	29.7	21.1	6	ไม่มี			0	
13												

กลุ่ม 1

	A	B	C	D	E	F	G	H	I	J	K	L
1	กลุ่ม 1											
2	0	44	154	73.5	44	22.1	15	ความดันโลหิตสูง			1	
3	0	59	165	72	38.6	23.8	15	เบาหวาน	ความดันโลหิตสูง		1	
4	0	40	155	74.3	46.8	21.1	17	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	1	
5	0	40	150	86.1	51.2	22.8	20	ความดันโลหิตสูง			1	
6	0	46	162	71.8	37.3	24.7	14	ความดันโลหิตสูง			1	
7	0	46	167	67.9	38.7	22.5	13	ความดันโลหิตสูง			1	
8	0	49	155	65.8	37.7	22.3	11	ความดันโลหิตสูง			1	
9	0	47	155	70	36	22	13	ความดันโลหิตสูง	เบาหวาน	ไขมันในเลือด	1	
10	0	48	155	56.4	41	19	11	ความดันโลหิตสูง	เบาหวาน		1	
11	0	45	155	55	40.2	17.7	11	ไม่มี			1	
12												

กลุ่ม 2

	A	B	C	D	E	F	G	H	I	J	K	L
1	กลุ่ม 2											
2	1	48	160	73.3	30.1	28.7	9	ไม่มี			2	
3	1	59	165	91.9	38.4	28.3	11	เบาหวาน			2	
4	1	57	175	82	27.8	32.7	10	ความดันโลหิตสูง	ไขมันในเลือดสูง		2	
5	1	58	165	67.8	30.6	25.6	9	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	2	
6	1	71	164	66.4	28	26.2	8	เบาหวาน			2	
7	1	49	162	57.3	20.4	24.8	5	เบาหวาน	ไทรอยด์		2	
8	1	67	165	65.3	26.1	26.2	8	เบาหวาน	ความดันโลหิตสูง		2	
9	1	52	162	67.5	31.5	25.1	10	เบาหวาน			2	
10	1	68	164	65	29.1	25	9	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	2	
11	1	57	168	68.7	19.3	31.2	6	ไม่มี			2	
12	1	57	164.5	87.7	38.6	30.1	17	เบาหวาน	ไขมันในเลือดสูง	เกาส์	2	
13	1	49	161	67	30.8	25.4	8	ไขมันในเลือดสูง			2	
14	1	56	160	76.5	34.9	27.8	12	ความดันโลหิตสูง			2	
15	1	51	164	68.8	29	27	8	ความดันโลหิตสูง	ไขมันในเลือดสูง		2	
16	1	42	185	76.5	22.2	33.1	6	ไม่มี			2	
17	1	44	170	67.5	25.8	27.9	7	ความดันโลหิตสูง			2	
18	1	45	175	75.5	24.8	31.6	9	เบาหวาน	ความดันโลหิตสูง	เกาส์	2	
19	1	49	175	71.6	18.9	32.5	5	ไม่มี			2	
20	1	55	163	88	39.5	30	15	ความดันโลหิตสูง	เบาหวาน	ไขมันในเลือด	2	
21												
22												

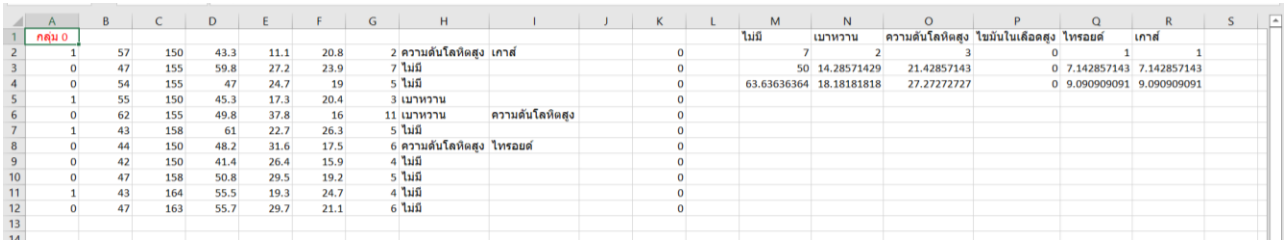
4. พิจารณาโรคที่พบบ่อยในผู้สูงอายุจากข้อมูลแต่ละกลุ่ม ดังนี้

กลุ่ม 0 จำนวนข้อมูล 11 ข้อมูล

คำนวณเปอร์เซ็นต์การเกิดโรคที่พบบ่อยในผู้สูงอายุต่อจำนวนคนทั้งหมดในกลุ่ม

โรคไม่ติดต่อที่พบบ่อย	โอกาสเกิดโรคไม่ติดต่อที่พบบ่อย (%)
ไม่มีโรค	63.6
ความดันโลหิตสูง	27.27
เบาหวาน	18.18
เกาส์	9.09
ไทรอยด์	9.09
ไขมันในเลือดสูง	0

จากตารางได้ว่า กลุ่มตัวอย่างกลุ่มนี้คาดการณ์ว่ามีโอกาสเกิดโรคความดันโลหิตสูงหรือโรคเบาหวานน้อยกว่า 30% และมีโอกาสที่จะเป็นโรคไขมันในเลือดสูงน้อยมาก



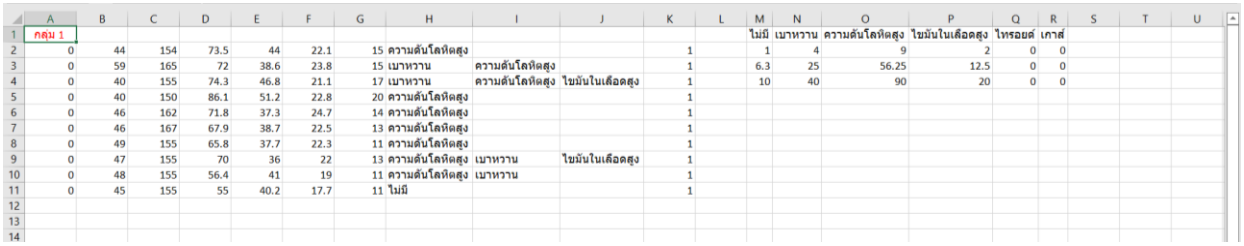
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	กลุ่ม 0												ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์	
2	1	57	150	43.3	11.1	20.8	2	ความดันโลหิตสูง	เกาส์		0		7	2	3	0	1	1	
3	0	47	155	59.8	27.2	23.9	7	ไม่มี			0		50	14.28571429	21.42857143	0	7.142857143	7.142857143	
4	0	54	155	47	24.7	19	5	ไม่มี			0		63.63636364	18.18181818	27.27272727	0	9.090909091	9.090909091	
5	1	55	150	45.3	17.3	20.4	3	เบาหวาน			0								
6	0	62	155	49.8	37.8	16	11	เบาหวาน	ความดันโลหิตสูง		0								
7	1	43	158	61	22.7	26.3	5	ไม่มี			0								
8	0	44	150	48.2	31.6	17.5	6	ความดันโลหิตสูง	ไทรอยด์		0								
9	0	42	150	41.4	26.4	15.9	4	ไม่มี			0								
10	0	47	158	50.8	29.5	19.2	5	ไม่มี			0								
11	1	43	164	55.5	19.3	24.7	4	ไม่มี			0								
12	0	47	163	55.7	29.7	21.1	6	ไม่มี			0								
13																			

กลุ่ม 1 จำนวนข้อมูล 10 ข้อมูล

คำนวณเปอร์เซ็นต์การเกิดโรคที่พบบ่อยในผู้สูงอายุต่อจำนวนคนทั้งหมดในกลุ่ม

โรคไม่ติดต่อที่พบบ่อย	โอกาสเกิดโรคไม่ติดต่อที่พบบ่อย (%)
ความดันโลหิตสูง	90
เบาหวาน	40
ไม่มีโรค	10
ไขมันในเลือดสูง	10
ไทรอยด์	0
เกาส์	0

จากตารางได้ว่า กลุ่มตัวอย่างกลุ่มนี้คาดการณ์ว่ามีโอกาสที่จะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานมากกว่า 30%



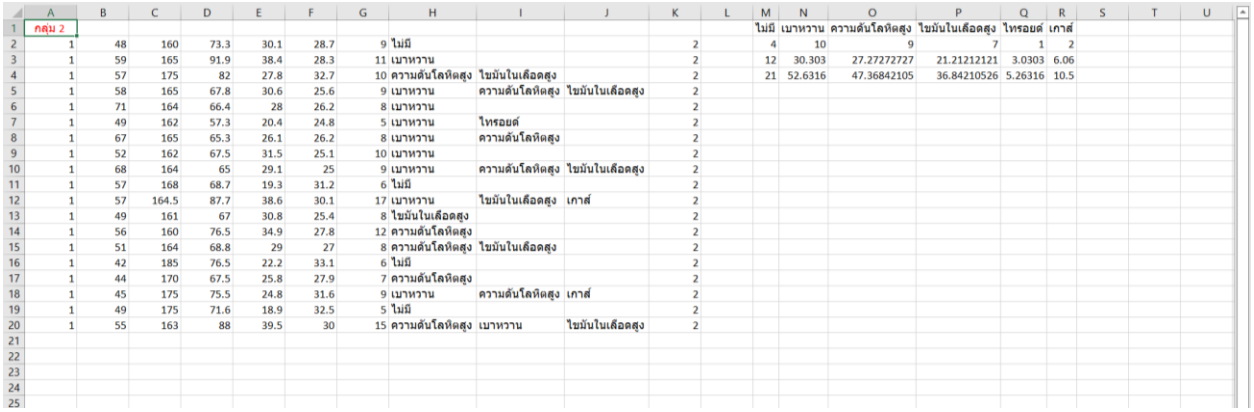
กลุ่ม 1												ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
0	44	154	73.5	44	22.1	15	ความดันโลหิตสูง			1		1	4	9	2	0	0
0	59	165	72	38.6	23.8	15	เบาหวาน	ความดันโลหิตสูง		1		6.3	25	56.25	12.5	0	0
0	40	155	74.3	46.8	21.1	17	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	1		10	40	90	20	0	0
0	40	150	86.1	51.2	22.8	20	ความดันโลหิตสูง			1							
0	46	162	71.8	37.3	24.7	14	ความดันโลหิตสูง			1							
0	46	167	67.9	38.7	22.5	13	ความดันโลหิตสูง			1							
0	49	155	65.8	37.7	22.3	11	ความดันโลหิตสูง			1							
0	47	155	70	36	22	13	ความดันโลหิตสูง	เบาหวาน	ไขมันในเลือดสูง	1							
0	48	155	56.4	41	19	11	ความดันโลหิตสูง	เบาหวาน		1							
0	45	155	55	40.2	17.7	11	ไม่มี			1							

กลุ่ม 2 จำนวนข้อมูล 19 ข้อมูล

คำนวณเปอร์เซ็นต์การเกิดโรคที่พบบ่อยในผู้สูงอายุต่อจำนวนคนทั้งหมดในกลุ่ม

โรคไม่ติดต่อที่พบบ่อย	โอกาสเกิดโรคไม่ติดต่อที่พบบ่อย (%)
เบาหวาน	52.63
ความดันโลหิตสูง	47.37
ไขมันในเลือดสูง	31.58
ไม่มีโรค	21.10
เกาส์	10.50
ไทรอยด์	3.13

จากตารางได้ว่า กลุ่มตัวอย่างกลุ่มนี้คาดการณ์ว่ามีโอกาสที่จะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานหรือโรคไขมันในเลือดสูงมากกว่า 30%



กลุ่ม 2												ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
1	48	160	73.3	30.1	28.7	9	ไม่มี			2		4	10	9	7	1	2
1	59	165	91.9	38.4	28.3	11	เบาหวาน			2		12	30.303	27.27272727	21.21212121	3.0303	6.06
1	57	175	82	27.8	32.7	10	ความดันโลหิตสูง	ไขมันในเลือดสูง		2		21	52.6316	47.36842105	36.84210526	5.26316	10.5
1	58	165	67.8	30.6	25.6	9	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	2							
1	71	164	66.4	28	26.2	8	เบาหวาน			2							
1	49	162	57.3	20.4	24.8	5	เบาหวาน	ไทรอยด์		2							
1	67	165	65.3	26.1	26.2	8	เบาหวาน	ความดันโลหิตสูง		2							
1	52	162	67.5	31.5	25.1	10	เบาหวาน			2							
1	68	164	65	29.1	25	9	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	2							
1	57	168	68.7	19.3	31.2	6	ไม่มี			2							
1	57	164.5	87.7	38.6	30.1	17	เบาหวาน	ไขมันในเลือดสูง	เกาส์	2							
1	49	161	67	30.8	25.4	8	ไขมันในเลือดสูง			2							
1	56	160	76.5	34.9	27.8	12	ความดันโลหิตสูง			2							
1	51	164	68.8	29	27	8	ความดันโลหิตสูง	ไขมันในเลือดสูง		2							
1	42	185	76.5	22.2	33.1	6	ไม่มี			2							
1	44	170	67.5	25.8	27.9	7	ความดันโลหิตสูง			2							
1	45	175	75.5	24.8	31.6	9	เบาหวาน	ความดันโลหิตสูง	เกาส์	2							
1	49	175	71.6	18.9	32.5	5	ไม่มี			2							
1	55	163	88	39.5	30	15	ความดันโลหิตสูง	เบาหวาน	ไขมันในเลือดสูง	2							

จะสังเกตได้ว่า จากการคำนวณจะสามารถคาดการณ์แนวโน้มโรคประจำกลุ่มจากเปอร์เซ็นต์สูงสุดได้ดังนี้

กลุ่ม 0 คาดการณ์ว่ามีโอกาสเกิดโรคความดันโลหิตสูงหรือโรคเบาหวานน้อยกว่า 30% และมีโอกาสที่จะเป็นโรคไขมันในเลือดสูงน้อยมาก

กลุ่ม 1 คาดการณ์ว่ามีโอกาสที่จะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานมากกว่า 30%

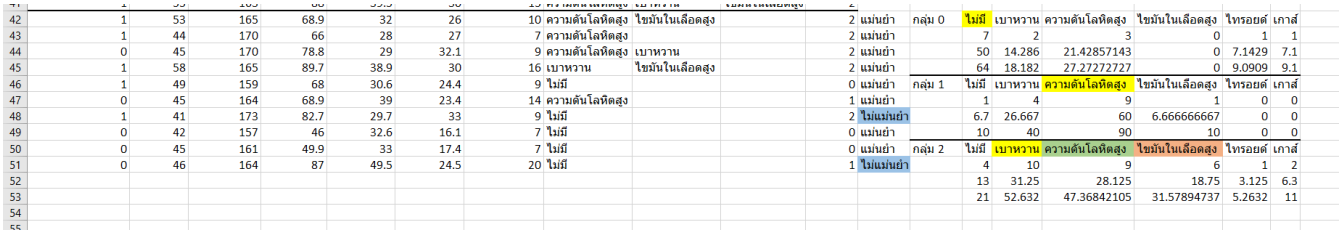
กลุ่ม 2 คาดการณ์ว่ามีโอกาสที่จะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานหรือโรคไขมันในเลือดสูงมากกว่า 30%

กลุ่ม 0	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
	7	2	3	0	1	1
	50	14.2857	21.42857143	0	7.14286	7.14
	63.6	18.1818	27.27272727	0	9.09091	9.09
กลุ่ม 1	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
	1	4	9	1	0	0
	6.67	26.6667	60	6.666666667	0	0
	10	40	90	10	0	0
กลุ่ม 2	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
	4	10	9	6	1	2
	12.5	31.25	28.125	18.75	3.125	6.25
	21.1	52.6316	47.36842105	31.57894737	5.26316	10.5

5. นำค่าที่พิจารณาได้จากแต่ละกลุ่มมาวิเคราะห์กับข้อมูลทดสอบ ได้ผลดังนี้

ข้อมูลทดสอบ	กลุ่ม	โรคจากข้อมูลจริง	โรคที่คาดการณ์	ความแม่นยำ
บุคคลที่ 1	2	ความดันโลหิตสูง, ไขมันในเลือดสูง	ความดันโลหิตสูง,เบาหวาน,ไขมันในเลือดสูงมากกว่า 30%	แม่นยำ
บุคคลที่ 2	2	ความดันโลหิตสูง	ความดันโลหิตสูง,เบาหวาน,ไขมันในเลือดสูงมากกว่า 30%	แม่นยำ
บุคคลที่ 3	2	ความดันโลหิตสูง, เบาหวาน	ความดันโลหิตสูง,เบาหวาน,ไขมันในเลือดสูงมากกว่า 30%	แม่นยำ

บุคคลที่ 4	2	เบาหวาน, ไขมันในเลือดสูง	ความดันโลหิตสูง,เบาหวาน,ไขมันในเลือดสูง มากกว่า 30%	แม่นยำ
บุคคลที่ 5	0	ไม่มี	ความดันโลหิตสูง,เบาหวาน น้อยกว่า 30%	แม่นยำ
บุคคลที่ 6	1	ความดันโลหิตสูง	ความดันโลหิตสูง,เบาหวาน มากกว่า 30%	แม่นยำ
บุคคลที่ 7	2	ไม่มี	ความดันโลหิตสูง,เบาหวาน,ไขมันในเลือดสูง มากกว่า 30%	ไม่แม่นยำ
บุคคลที่ 8	0	ไม่มี	ความดันโลหิตสูง,เบาหวาน น้อยกว่า 30%	แม่นยำ
บุคคลที่ 9	0	ไม่มี	ความดันโลหิตสูง,เบาหวาน น้อยกว่า 30%	แม่นยำ
บุคคลที่ 10	1	ไม่มี	ความดันโลหิตสูง,เบาหวาน มากกว่า 30%	ไม่แม่นยำ



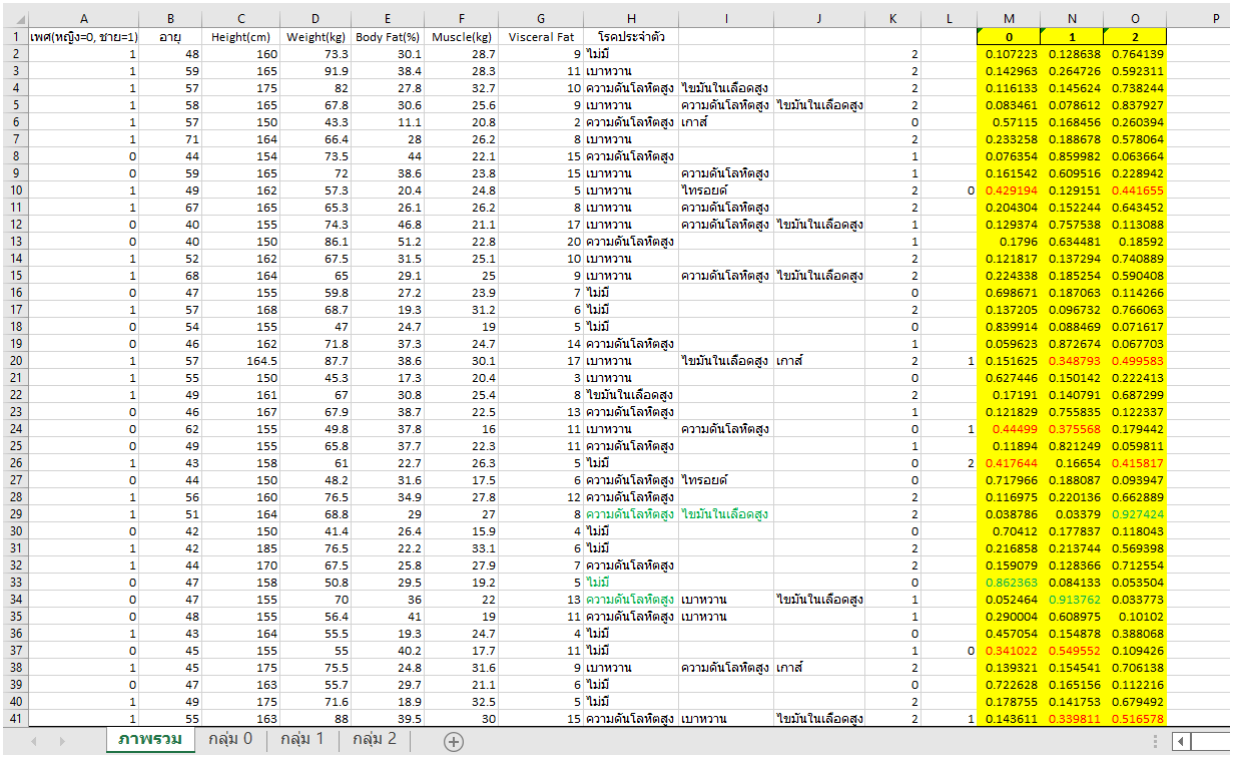
42	1	53	165	68.9	32	26	10	ความดันโลหิตสูง	ไขมันในเลือดสูง	2	แม่นยำ	กลุ่ม 0	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาต์
43	1	44	170	66	28	27	7	ความดันโลหิตสูง		2	แม่นยำ		7	2	3	0	1	1
44	0	45	170	78.8	29	32.1	9	ความดันโลหิตสูง	เบาหวาน	2	แม่นยำ		50	14.286	21.42857143	0	7.1429	7.1
45	1	58	165	89.7	38.9	30	16	เบาหวาน	ไขมันในเลือดสูง	2	แม่นยำ		64	18.182	27.27272727	0	9.0909	9.1
46	1	49	159	68	30.6	24.4	9	ไม่มี		0	แม่นยำ	กลุ่ม 1	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาต์
47	0	45	164	68.9	39	23.4	14	ความดันโลหิตสูง		1	แม่นยำ		1	4	9	1	0	0
48	1	41	173	82.7	29.7	33	9	ไม่มี		2	ไม่แม่นยำ		6.7	26.667	60	6.666666667	0	0
49	0	42	157	46	32.6	16.1	7	ไม่มี		0	แม่นยำ		10	40	90	10	0	0
50	0	45	161	49.9	33	17.4	7	ไม่มี		0	แม่นยำ	กลุ่ม 2	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาต์
51	0	46	164	87	49.5	24.5	20	ไม่มี		1	ไม่แม่นยำ		4	10	9	6	1	2
52													13	31.25	28.125	18.75	3.125	6.3
53													21	52.632	47.36842105	31.57894737	5.2632	11
54																		

จากผลสรุปความแม่นยำของแบบจำลอง พบว่ามีความแม่นยำในอัตราส่วน 8 ต่อ 10 ข้อมูลและจากการสังเกตข้อมูลที่ไม่แม่นยำ คือข้อมูลบุคคลที่ 7 และบุคคลที่ 10 ซึ่งจากการเก็บข้อมูลบุคคลดังกล่าวไม่มีโรคประจำตัว ทำให้เราสามารถสันนิษฐานเบื้องต้นได้ว่าบุคคลที่ 7 มีความเป็นไปได้ที่จะเกิด “โรคความดันโลหิตสูงหรือโรคเบาหวานหรือโรคไขมันในเลือดสูงมากกว่า 30%” และบุคคลที่ 10 มีความเป็นไปได้ที่จะเกิด “โรคความดันโลหิตสูงหรือโรคเบาหวานมากกว่า 30%” โดยที่ยังไม่ทราบ ดังนั้นสามารถให้คำแนะนำบุคคลดังกล่าวเข้าพบแพทย์หรือปรับปรุงพฤติกรรมของตนเอง เพื่อป้องกันการเกิดโรคได้

3.6 วิเคราะห์โรคไม่ติดต่อที่พบบ่อยเทียบกับผลลัพธ์ที่ได้จากแบบจำลอง Fuzzy C-Means แบบ Cluster Intersection

ในหัวข้อนี้จะวิเคราะห์การแบ่งกลุ่ม โดยกำหนดให้บางข้อมูลอยู่ได้มากกว่าหนึ่งกลุ่ม ตามคุณสมบัติสำคัญของวิธีแบ่งกลุ่มแบบ Fuzzy C-Means โดยพิจารณาจากค่าความเป็นสมาชิก (Membership Function, μ)

1. จากการคำนวณผ่านแบบจำลองด้วยโปรแกรมภาษา Python เพื่อหาค่าความเป็นสมาชิกได้ผลลัพธ์ ดังรูป



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	เพศ(หญิง=0, ชาย=1)	อายุ	Height(cm)	Weight(kg)	Body Fat(%)	Muscle(kg)	Visceral Fat	โรคประจำตัว					0	1	2	
2	1	48	160	73.3	30.1	28.7	9	ไม่มี			2		0.107223	0.128638	0.764139	
3	1	59	165	91.9	38.4	28.3	11	เบาหวาน			2		0.142963	0.264726	0.592311	
4	1	57	175	82	27.8	32.7	10	ความดันโลหิตสูง	ไขมันในเลือดสูง		2		0.116133	0.145624	0.738244	
5	1	58	165	67.8	30.6	25.6	9	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	2		0.083461	0.078612	0.837927	
6	1	57	150	43.3	11.1	20.8	2	ความดันโลหิตสูง	เกาต์		0		0.57115	0.168456	0.260394	
7	1	71	164	66.4	28	26.2	8	เบาหวาน			2		0.233258	0.188678	0.578064	
8	0	44	154	73.5	44	22.1	15	ความดันโลหิตสูง			1		0.076354	0.859982	0.063664	
9	0	59	165	72	38.6	23.8	15	เบาหวาน	ความดันโลหิตสูง		1		0.161542	0.609516	0.228942	
10	1	49	162	57.3	20.4	24.8	5	เบาหวาน	ไทรอยด์		2	0	0.429194	0.129151	0.441655	
11	1	67	165	65.3	26.1	26.2	8	เบาหวาน	ความดันโลหิตสูง		2		0.204304	0.152244	0.643452	
12	0	40	155	74.3	46.8	21.1	17	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	1		0.129374	0.757538	0.113088	
13	0	40	150	86.1	51.2	22.8	20	ความดันโลหิตสูง			1		0.1796	0.634481	0.18592	
14	1	52	162	67.5	31.5	25.1	10	เบาหวาน			2		0.121817	0.137294	0.740889	
15	1	68	164	65	29.1	25	9	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	2		0.224338	0.185254	0.590408	
16	0	47	155	59.8	27.2	23.9	7	ไม่มี			0		0.698671	0.187063	0.114266	
17	1	57	168	68.7	19.3	31.2	6	ไม่มี			2		0.137205	0.096732	0.766063	
18	0	54	155	47	24.7	19	5	ไม่มี			0		0.839914	0.088469	0.071617	
19	0	46	162	71.8	37.3	24.7	14	ความดันโลหิตสูง			1		0.059623	0.872674	0.067703	
20	1	57	164.5	87.7	38.6	30.1	17	เบาหวาน	ไขมันในเลือดสูง	เกาต์	2	1	0.151625	0.348793	0.499583	
21	1	55	150	45.3	17.3	20.4	3	เบาหวาน			0		0.627446	0.150142	0.222413	
22	1	49	161	67	30.8	25.4	8	ไขมันในเลือดสูง			2		0.17191	0.140791	0.687299	
23	0	46	167	67.9	38.7	22.5	13	ความดันโลหิตสูง			1		0.121829	0.755835	0.122337	
24	0	62	155	49.8	37.8	16	11	เบาหวาน	ความดันโลหิตสูง		0	1	0.44499	0.375568	0.179442	
25	0	49	155	65.8	37.7	22.3	11	ความดันโลหิตสูง			1		0.11894	0.821249	0.059811	
26	1	43	158	61	22.7	26.3	5	ไม่มี			0	2	0.417644	0.16654	0.415817	
27	0	44	150	48.2	31.6	17.5	6	ความดันโลหิตสูง	ไทรอยด์		0		0.717966	0.188087	0.093947	
28	1	56	160	76.5	34.9	27.8	12	ความดันโลหิตสูง			2		0.116975	0.220136	0.662889	
29	1	51	164	68.8	29	27	8	ความดันโลหิตสูง	ไขมันในเลือดสูง		2		0.038786	0.03379	0.927424	
30	0	42	150	41.4	26.4	15.9	4	ไม่มี			0		0.70412	0.177837	0.118043	
31	1	42	185	76.5	22.2	33.1	6	ไม่มี			2		0.216858	0.213744	0.569398	
32	1	44	170	67.5	25.8	27.9	7	ความดันโลหิตสูง			2		0.159079	0.128366	0.712554	
33	0	47	158	50.8	29.5	19.2	5	ไม่มี			0		0.862363	0.084133	0.053504	
34	0	47	155	70	36	22	13	ความดันโลหิตสูง	เบาหวาน	ไขมันในเลือดสูง	1		0.052464	0.913762	0.033773	
35	0	48	155	56.4	41	19	11	ความดันโลหิตสูง	เบาหวาน		1		0.290004	0.608975	0.10102	
36	1	43	164	55.5	19.3	24.7	4	ไม่มี			0		0.457054	0.154878	0.388068	
37	0	45	155	55	40.2	17.7	11	ไม่มี			1	0	0.341022	0.549552	0.109426	
38	1	45	175	75.5	24.8	31.6	9	เบาหวาน	ความดันโลหิตสูง	เกาต์	2		0.139321	0.154541	0.706138	
39	0	47	163	55.7	29.7	21.1	6	ไม่มี			0		0.722628	0.165156	0.112216	
40	1	49	175	71.6	18.9	32.5	5	ไม่มี			2		0.178755	0.141753	0.679492	
41	1	55	163	88	39.5	30	15	ความดันโลหิตสูง	เบาหวาน	ไขมันในเลือดสูง	2	1	0.143611	0.339811	0.516578	

ภาพรวม | กลุ่ม 0 | กลุ่ม 1 | กลุ่ม 2

เห็นได้ว่า มีข้อมูลสำหรับการเรียนรู้ 6 ข้อมูล ที่สามารถอยู่ได้มากกว่า 1 กลุ่ม โดยพิจารณาจากอัตราส่วนเทียบกับค่าความเป็นสมาชิกในกลุ่มหลักที่มากกว่า 60% ดังนี้

ข้อมูลสำหรับการเรียนรู้	กลุ่มหลัก	กลุ่มรอง
ข้อมูลบุคคลที่ 9	2 (มีโอกาส 44.17%)	0 (มีโอกาส 42.92%)
ข้อมูลบุคคลที่ 19	2 (มีโอกาส 49.96%)	1 (มีโอกาส 34.88%)
ข้อมูลบุคคลที่ 23	0 (มีโอกาส 44.50%)	1 (มีโอกาส 37.56%)
ข้อมูลบุคคลที่ 25	0 (มีโอกาส 41.76%)	2 (มีโอกาส 41.58%)
ข้อมูลบุคคลที่ 36	1 (มีโอกาส 54.96%)	0 (มีโอกาส 34.10%)
ข้อมูลบุคคลที่ 40	2 (มีโอกาส 51.66%)	1 (มีโอกาส 33.98%)

2. คำนวณเปอร์เซ็นต์การเกิดโรคไม่ติดต่อที่พบบ่อยต่อจำนวนคนทั้งหมดในแต่ละกลุ่ม ได้ผลลัพธ์ดังรูป

Cluster Intersection	กลุ่ม 0	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
		8	3	3	0	2	1
		6.1305395	2.2989523	2.29895232	0	1.53263	0.77
		61.538462	23.076923	23.07692308	0	15.3846	7.69
	กลุ่ม 1	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
		1	7	11	3	0	1
		0.7663174	5.3642221	8.429491841	2.29895232	0	0.77
		7.6923077	53.846154	84.61538462	23.07692308	0	7.69
	กลุ่ม 2	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
		5	10	9	6	1	2
		3.8315872	7.6631744	6.896856961	4.59790464	0.76632	1.53
		25	50	45	30	5	10

จะเห็นได้ว่าเปอร์เซ็นต์การเกิดโรคไม่แตกต่างจากการแบ่งกลุ่มในหัวข้อ 3.4 แต่ผลลัพธ์ของการแบ่งกลุ่มแบบ Cluster Intersection จะมีความยืดหยุ่นกว่า

บทที่ 4

ผลการศึกษา

4.1 สรุปผลการศึกษา

จากการศึกษาความสัมพันธ์ระหว่างตัวแปรต้นทั้ง 7 ตัว คือ เพศ อายุ ส่วนสูง (Height) น้ำหนัก (Weight) ไขมันในร่างกาย (Body Fat) กล้ามเนื้อ (Muscle) และไขมันในช่องท้อง (Visceral Fat) ตามค่าที่วัดได้จากเครื่องชั่งอัจฉริยะ InBody Dial Body Composition Analyzer และข้อมูลแบบฟอร์มสำรวจ เทียบกับผลลัพธ์ คือ โรคไม่ติดต่อที่พบบ่อย ด้วยวิธีการจัดกลุ่มแบบ Fuzzy C-Means แสดงให้เห็นว่า ข้อมูลถูกแบ่งออกเป็น 3 กลุ่ม ดังนี้

กลุ่ม 0 คาดการณ์ว่ามีโอกาสเกิดโรคความดันโลหิตสูงหรือโรคเบาหวานน้อยกว่า 30% และมีโอกาสที่จะเป็นโรคไขมันในเลือดสูงน้อยมาก

กลุ่ม 1 คาดการณ์ว่ามีโอกาสที่จะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานมากกว่า 30%

กลุ่ม 2 คาดการณ์ว่ามีโอกาสที่จะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานหรือโรคไขมันในเลือดสูงมากกว่า 30%

หลักจากทดสอบแบบจำลองโดยใช้ข้อมูลทดสอบ เพื่อตรวจสอบความแม่นยำ พบว่ามีความแม่นยำในอัตราส่วน 8 ต่อ 10 ข้อมูลและจากการสังเกตข้อมูลที่ไม่แม่นยำ คือข้อมูลบุคคลที่ 7 และบุคคลที่ 10 ซึ่งจากการเก็บข้อมูลบุคคลดังกล่าวไม่มีโรคประจำตัว ทำให้เราสามารถสันนิษฐานเบื้องต้นได้ว่าบุคคลที่ 7 มีความเป็นไปได้ที่จะเกิด “โรคความดันโลหิตสูงหรือโรคเบาหวานหรือโรคไขมันในเลือดสูงมากกว่า 30%” และบุคคลที่ 10 มีความเป็นไปได้ที่จะเกิด “โรคความดันโลหิตสูงหรือโรคเบาหวานมากกว่า 30%” โดยที่ยังไม่ทราบ ดังนั้นในเบื้องต้นสามารถให้คำแนะนำบุคคลดังกล่าวเข้าพบแพทย์หรือปรับปรุงพฤติกรรมของตนเอง เพื่อป้องกันการเกิดโรคติดต่อที่พบบ่อยได้

4.2 อภิปรายผลการศึกษา

จากผลการดำเนินงานค้นคว้าอิสระข้างต้น เราสามารถตั้งข้อสังเกตเพิ่มเติม ดังนี้

4.2.1 โรคเกาส์และโรคไทรอยด์มีผลในการดำเนินงานครั้งนี้น้อย เนื่องจากข้อมูลไม่เพียงพอ

4.2.2 คาดว่าข้อสรุปจะละเอียดมากขึ้น หากมีข้อมูลสำหรับการเรียนรู้มากขึ้น

4.2.3 ตัวแปรต้นที่นำมาพิจารณาหากมีความสัมพันธ์กับโรคไม่ติดต่อที่พบบ่อยตามหลักการแพทย์อาจส่งผลให้ผลลัพธ์ที่ได้แม่นยำมากขึ้น

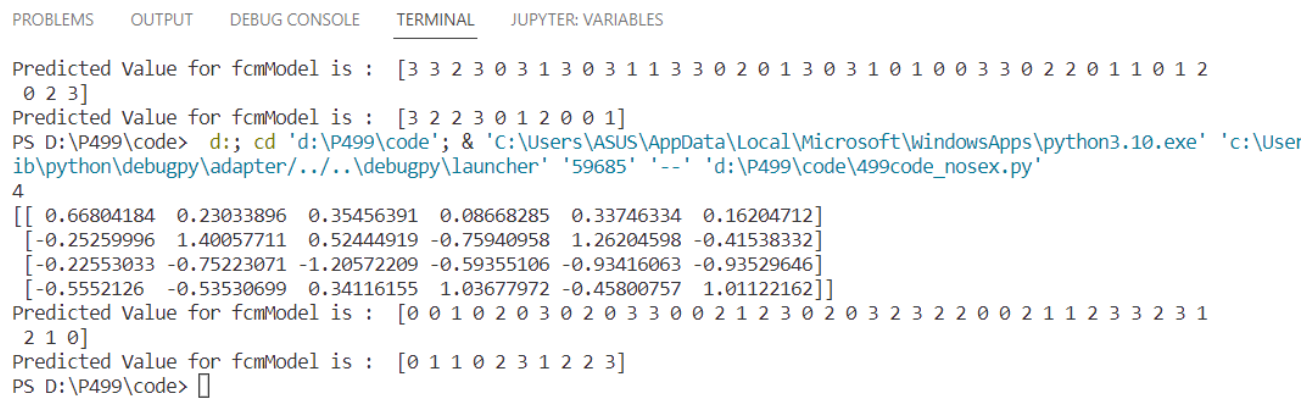
4.2.4 ข้อดีของวิธีแบ่งกลุ่มแบบ Fuzzy C-Means ได้ผลลัพธ์ในการระบุเปอร์เซ็นต์การเกิดโรคมีความยืดหยุ่นกว่าวิธีการแบ่งกลุ่มแบบ K-Means และยังสามารถระบุข้อมูลสำหรับการเรียนรู้ว่าจัดอยู่ในกลุ่มใด คิดเป็นเปอร์เซ็นต์

4.2.5 ค่าความเป็นสมาชิกที่สูงในแต่ละกลุ่มสอดคล้องกับโรคที่ระบุในแต่ละกลุ่ม ดังนี้ กลุ่ม 0 ข้อมูลบุคคลที่ 32 มีค่าความเป็นสมาชิกมากที่สุดเท่ากับ 0.862 คาดการณ์ว่ามีโอกาสเกิดโรคความดันโลหิตสูงหรือโรคเบาหวานน้อยกว่า 30% และมีโอกาสที่จะเป็นโรคไขมันในเลือดสูงน้อยมาก, กลุ่ม 1 ข้อมูลบุคคลที่ 33 มีค่าความเป็นสมาชิกมากที่สุดเท่ากับ 0.914 คาดการณ์ว่ามีโอกาสที่จะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานมากกว่า 30%, กลุ่ม 2 ข้อมูลบุคคลที่ 28 มีค่าความเป็นสมาชิกมากที่สุดเท่ากับ 0.927 คาดการณ์ว่ามีโอกาสที่จะเป็นโรคความดันโลหิตสูงหรือโรคเบาหวานหรือโรคไขมันในเลือดสูงมากกว่า 30%

4.2.6 ตัวแปรต้นที่นำมาอาจมีความผิดพลาดรบกวน (noise) หากได้กรองข้อมูลในเบื้องต้นก่อน อาจส่งผลให้ผลลัพธ์ที่ได้แม่นยำมากขึ้น

4.2.7 จากการพิจารณาข้อมูลแต่ละกลุ่ม สังเกตพบว่ากลุ่ม 1 เป็นเพศหญิงทั้งหมดและกลุ่ม 2 เป็นเพศชายทั้งหมด เราจึงสร้างแบบจำลองใหม่อีกชุด โดยไม่นำตัวแปร “เพศ” มาใช้ในการคำนวณ เนื่องจากเห็นว่าค่าตัวแปรนี้แบ่งกลุ่มชัดเจนตามผลลัพธ์อยู่แล้ว ดังนั้นจะตรวจสอบว่าการใช้ตัวแปรเพียง 6 ตัว คือ A, H, W, BF, M, VF สามารถแบ่งกลุ่มโรคไม่ติดต่อที่พบบ่อยและกลุ่มเพศของกลุ่มบุคคลได้ด้วยหรือไม่

ซึ่งวิธีการเตรียมข้อมูลทำเหมือนแบบจำลองชุดแรกแต่ตัดตัวแปร เพศ หรือ S ออก พบว่าสำหรับการแยกกลุ่มข้อมูลของโรคไม่ติดต่อที่พบบ่อยได้จำนวนกลุ่มที่เหมาะสมเพิ่มขึ้นเป็น 4 กลุ่ม ดังรูป



42	53	165	68.9	32	26	10	ความดันโลหิตสูง	ไขมันในเลือดสูง	0	แม่นยำ	กลุ่ม 0	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
43	44	170	66	28	27	7	ความดันโลหิตสูง		1	แม่นยำ		1	9	7	5	0	1
44	45	170	78.8	29	32.1	9	ความดันโลหิตสูง	เบาหวาน	1	แม่นยำ		4.34782609	39.1304348	30.43478261	21.73913043	0	4.34782609
45	58	165	89.7	38.9	30	16	เบาหวาน	ไขมันในเลือดสูง	0	แม่นยำ		7.69230769	69.2307692	53.84615385	38.46153846	0	7.69230769
46	49	159	68	30.6	24.4	9	ไม่มี		3	แม่นยำ	กลุ่ม 1	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
47	45	164	68.9	39	23.4	14	ความดันโลหิตสูง		2	แม่นยำ		3	1	3	1	0	1
48	41	173	82.7	29.7	33	9	ไม่มี		1	แม่นยำ		33.3333333	11.1111111	33.33333333	11.11111111	0	11.1111111
49	42	157	46	32.6	16.1	7	ไม่มี		3	แม่นยำ		50	16.6666667	50	16.66666667	0	16.6666667
50	45	161	49.9	33	17.4	7	ไม่มี		3	แม่นยำ	กลุ่ม 2	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
51	46	164	87	49.5	24.5	20	ไม่มี		2	ไม่แม่นยำ		1	3	8	2	0	0
52												7.14285714	21.4285714	57.14285714	14.28571429	0	0
53												11.1111111	33.3333333	88.88888889	22.22222222	0	0
54											กลุ่ม 3	ไม่มี	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	ไทรอยด์	เกาส์
55												7	3	3	0	2	1
56												43.75	18.75	18.75	0	12.5	6.25
57												58.3333333	25	25	0	16.667	8.33333333

ซึ่งความแม่นยำเพิ่มขึ้นเป็น 9 ต่อ 10 ข้อมูลอย่างมีนัยสำคัญ

สำหรับการแยกข้อมูลของเพศ แสดงดังรูป

กลุ่ม 0

	A	B	C	D	E	F	G	H	I	J	K
1	กลุ่ม 0										
2	1	48	160	73.3	30.1	28.7	9	ไม่มี			0
3	1	59	165	91.9	38.4	28.3	11	เบาหวาน			0
4	1	58	165	67.8	30.6	25.6	9	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	0
5	1	71	164	66.4	28	26.2	8	เบาหวาน			0
6	0	59	165	72	38.6	23.8	15	เบาหวาน	ความดันโลหิตสูง		0
7	1	67	165	65.3	26.1	26.2	8	เบาหวาน	ความดันโลหิตสูง		0
8	1	52	162	67.5	31.5	25.1	10	เบาหวาน			0
9	1	68	164	65	29.1	25	9	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	0
10	1	57	164.5	87.7	38.6	30.1	17	เบาหวาน	ไขมันในเลือดสูง	เกาส์	0
11	1	49	161	67	30.8	25.4	8	ไขมันในเลือดสูง			0
12	1	56	160	76.5	34.9	27.8	12	ความดันโลหิตสูง			0
13	1	51	164	68.8	29	27	8	ความดันโลหิตสูง	ไขมันในเลือดสูง		0
14	1	55	163	88	39.5	30	15	ความดันโลหิตสูง	เบาหวาน	ไขมันในเลือด	0
15											
16											

กลุ่ม 1

	A	B	C	D	E	F	G	H	I	J	K
1	กลุ่ม 1										
2	1	57	175	82	27.8	32.7	10	ความดันโลหิตสูง	ไขมันในเลือดสูง		1
3	1	57	168	68.7	19.3	31.2	6	ไม่มี			1
4	1	42	185	76.5	22.2	33.1	6	ไม่มี			1
5	1	44	170	67.5	25.8	27.9	7	ความดันโลหิตสูง			1
6	1	45	175	75.5	24.8	31.6	9	เบาหวาน	ความดันโลหิตสูง	เกาส์	1
7	1	49	175	71.6	18.9	32.5	5	ไม่มี			1
8											
9											

กลุ่ม 2

	A	B	C	D	E	F	G	H	I	J	K	L
1	กลุ่ม 2											
2	0	44	154	73.5	44	22.1	15	ความดันโลหิตสูง			2	
3	0	40	155	74.3	46.8	21.1	17	เบาหวาน	ความดันโลหิตสูง	ไขมันในเลือดสูง	2	
4	0	40	150	86.1	51.2	22.8	20	ความดันโลหิตสูง			2	
5	0	46	162	71.8	37.3	24.7	14	ความดันโลหิตสูง			2	
6	0	46	167	67.9	38.7	22.5	13	ความดันโลหิตสูง			2	
7	0	49	155	65.8	37.7	22.3	11	ความดันโลหิตสูง			2	
8	0	47	155	70	36	22	13	ความดันโลหิตสูง	เบาหวาน	ไขมันในเลือดสูง	2	
9	0	48	155	56.4	41	19	11	ความดันโลหิตสูง	เบาหวาน		2	
10	0	45	155	55	40.2	17.7	11	ไม่มี			2	
11												
12												

กลุ่ม 3

	A	B	C	D	E	F	G	H	I	J	K	L
1	กลุ่ม 3											
2	1	57	150	43.3	11.1	20.8	2	ความดันโลหิตสูง	เกาส์		3	
3	1	49	162	57.3	20.4	24.8	5	เบาหวาน	ไทรอยด์		3	
4	0	47	155	59.8	27.2	23.9	7	ไม่มี			3	
5	0	54	155	47	24.7	19	5	ไม่มี			3	
6	1	55	150	45.3	17.3	20.4	3	เบาหวาน			3	
7	0	62	155	49.8	37.8	16	11	เบาหวาน	ความดันโลหิตสูง		3	
8	1	43	158	61	22.7	26.3	5	ไม่มี			3	
9	0	44	150	48.2	31.6	17.5	6	ความดันโลหิตสูง	ไทรอยด์		3	
10	0	42	150	41.4	26.4	15.9	4	ไม่มี			3	
11	0	47	158	50.8	29.5	19.2	5	ไม่มี			3	
12	1	43	164	55.5	19.3	24.7	4	ไม่มี			3	
13	0	47	163	55.7	29.7	21.1	6	ไม่มี			3	
14												
15												

สรุปได้ว่า กลุ่ม 0 และกลุ่ม 1 สามารถระบุได้ว่าเป็นเพศชาย กลุ่ม 2 เป็นเพศหญิง ส่วนกลุ่ม 3 สรุปไม่ได้ สันนิษฐานว่าสอดคล้องกับสรีระมาตรฐานของแต่ละเพศ

เอกสารอ้างอิง

[1] นิตยา ณ เชียงใหม่, ปฤษณา กลับอุดม, “Math 112 แคลคูลัส 2”, ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่, 2552

[2] วัชรพงษ์ อนรรฆเมธี, “เอกสารประกอบการสอนกระบวนวิชา แคลคูลัสขั้นสูง”, ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร, 2564

[3] Calculus for Engineering II, ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่, 2554

[4] Danielsson Per-Erik. “Euclidean distance mapping”, Computer Graphics and Image Processing, 14(3), pp. 227 - 248. 1980

[5] Francios Chollet. “Deep Learning with Python”, Second Edition, 2021

[6] ปริญญา สงวนสัตย์, “Artificial Intelligence with Machine Learning”, Python Edition, 2562

[7] Weerasak Thachai. “การหาจำนวน k ที่เหมาะสมที่สุดด้วยวิธี Elbow Method”, EspressoFX Notebook, 2017

[8] Madson Luiz Dantas Dias, fuzzy c-means: An implementation of Fuzzy C-means clustering algorithm., Zenodo, 2019

ภาคผนวก

ภาคผนวก ก

ประมวลภาพการเก็บรวบรวมข้อมูลสำหรับงานค้นคว้าอิสระนี้ ณ โรงพยาบาลหริภุญชัย ราม จังหวัดลำพูน

ประมวลภาพการเก็บรวบรวมข้อมูลสำหรับงานค้นคว้าอิสระนี้ ณ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

ภาคผนวก ข

โปรแกรมภาษา Python สำหรับการจัดกลุ่มข้อมูลแบบฟัซซีซีมีนจากเครื่องชั่งน้ำหนักอัจฉริยะเพื่อวิเคราะห์โรคที่พบบ่อยในผู้สูงอายุ

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from fcmeans import FCM
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
from kneed import KneeLocator
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

data = pd.read_csv('D:\P499\code\data\datac.csv')
data = pd.DataFrame(data,columns=['S','A','H','W','BF','M','VF'])

#Spliting Data To X_train and X-test
X_train = data.iloc[:40,:]
X_test = data.iloc[40:,:]
#print(X_train.shape)
#print(X_test.shape)

#Scaling Data
scalarModel = StandardScaler()
X_train = scalarModel.fit_transform(X_train)
X_test = scalarModel.fit_transform(X_test)
#print(X_train)
#print(X_test)

#Choosing the Appropriate Number of Clusters
# A list holds the SSE values for each k
sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(X_train)
    sse.append(kmeans.inertia_)

plt.style.use("fivethirtyeight")
plt.plot(range(1, 11), sse)
plt.xticks(range(1, 11))
plt.xlabel("Number of Clusters")
plt.ylabel("SSE")
```

```
plt.close()

kl = KneeLocator(
    range(1, 11), sse, curve="convex", direction="decreasing"
)

print(kl.elbow)

#Apply the algorithm Fuzzy c Means
fcmModel = FCM(n_clusters =kl.elbow)
fcmModel.fit(X_train)
center = fcmModel.centers
print(center)

#Calculating Prediction
ptrain = fcmModel.predict(X_train)
ptest = fcmModel.predict(X_test)
print('Predicted Value for fcmModel is : ' , ptrain)
print('Predicted Value for fcmModel is : ' , ptest)


#Calculating Membership Function
u = pd.DataFrame(fcmModel.u,columns=['0','1','2'])
writer = pd.ExcelWriter('solutions.xlsx')
u.to_excel(writer, sheet_name='Sheet1')
writer.save()
```

ภาคผนวก ค

เอกสารขอความอนุเคราะห์ข้อมูลเพื่อใช้ในกระบวนวิชาค้นคว้าอิสระ (206499)

ที่ อว ๘๓๙๓(๑๓.๑)/๖๙๘ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
มหาวิทยาลัยเชียงใหม่ ๕๐๒๐๐

๙ สิงหาคม ๒๕๖๕

เรื่อง ขอความอนุเคราะห์ข้อมูลเพื่อใช้ในกระบวนวิชาค้นคว้าอิสระ (๒๐๖๔๙๙)

เรียน ผู้อำนวยการโรงพยาบาลหริภุญชัย ราม

ด้วยนางสาวปิยธิดา นวลเหลือ รหัสประจำตัวนักศึกษา ๖๒๐๕๑๐๕๑๑ นักศึกษาชั้นปีที่ ๔ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ได้ทำวิจัยในกระบวนวิชาค้นคว้าอิสระ (๒๐๖๔๙๙) ในหัวข้อ"การจัดกลุ่มข้อมูลแบบฟัซซีซีมีนจากเครื่องชั่งน้ำหนักอัจฉริยะเพื่อวิเคราะห์โรคที่พบบ่อยในผู้สูงอายุ (Fuzzy C-Means Clustering Based on Body Composition Scale for Analysis of Common Elderly Illnesses)" และมีความประสงค์จะขอความอนุเคราะห์ข้อมูลจากโรงพยาบาลหริภุญชัย ราม เพื่อศึกษาเพิ่มเติมสำหรับกระบวนวิชาค้นคว้าอิสระ (๒๐๖๔๙๙) โดยมีผู้ช่วยศาสตราจารย์ ดร.ณัฐวัชร์ สนธิชัย เป็นอาจารย์ที่ปรึกษา นั้น

ในการนี้ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ จึงใคร่ขอความอนุเคราะห์ข้อมูลกับทางโรงพยาบาลหริภุญชัย ราม ในการเปิดเผยข้อมูลบางส่วนของโรงพยาบาลหริภุญชัย ราม ดังนี้

ขั้นตอนการเก็บข้อมูล

๑. สอบถามชื่อ นามสกุล เบอร์โทร โรคประจำตัวและส่วนสูงของผู้ให้ข้อมูล
๒. วัดค่าร่างกายจากเครื่องชั่ง Body Composition Scale (ผู้จัดเก็บข้อมูลเตรียมมา)
๓. ผู้จัดเก็บข้อมูลบันทึกข้อมูลลงในแบบฟอร์มเก็บข้อมูล
๔. ให้ผู้ให้ข้อมูลเซ็นอนุญาตให้นำข้อมูลไปใช้

ทั้งนี้ข้อมูลเหล่านี้ใช้ในกระบวนวิชาค้นคว้าอิสระ (๒๐๖๔๙๙) เท่านั้น และจะเก็บเป็นความลับของผู้จัดเก็บข้อมูล โดยจะไม่เผยแพร่เป็นอันขาด ทั้งนี้ภาควิชาคณิตศาสตร์เห็นว่าการค้นคว้าอิสระดังกล่าว มีประโยชน์และเป็นการประยุกต์ความรู้ทางคณิตศาสตร์กับปัญหาในสถานการณ์จริง จึงขอความอนุเคราะห์ให้นักศึกษาได้เก็บข้อมูลดังกล่าว ในวันเสาร์ที่ ๒๐ สิงหาคม ๒๕๖๕ เวลา ๙.๐๐-๑๒.๐๐ น. จำนวน ๕๐ ข้อมูล พร้อมนี้ได้แนบแบบฟอร์มเก็บข้อมูลการค้นคว้าอิสระและจำนงลงทะเบียนกระบวนวิชา ๒๐๖๔๙๙ เพื่อประกอบการพิจารณา

จึงเรียนมาเพื่อโปรดพิจารณาให้ความอนุเคราะห์ จักขอบพระคุณยิ่ง

(รองศาสตราจารย์ ดร.ณัฐกร สุคันธมาลา)
หัวหน้าภาควิชาคณิตศาสตร์

ภาควิชาคณิตศาสตร์
โทรศัพท์ (๐๕๓) ๙๔๓๓๒๖ ต่อ ๑๐๔

ภาคผนวก ง

โปสเตอร์กระบวนวิชาค้นคว้าอิสระ (206499)

Fuzzy C-Means Clustering Based on Body Composition Scale for Analysis of Non-Communicable Diseases

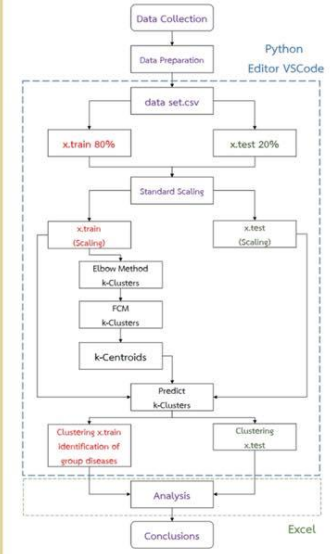
Piyathida Nuanlua Code: 620510511
Advisor: Asst. Prof. Dr. Nuttawat Sontichai
Department of Mathematics, Faculty of Science, Chiang Mai University

HARIPHUNCHAI RAM HOSPITAL

Abstract

The objective of this independent study is to identify the population affected by non-communicable diseases. Utilizing Fuzzy C-Means clustering and the Body Composition Scale, its measurable value is determined. Additionally, cluster sampling is utilized to observe the trend of non-communicable diseases. In addition, research revealed that data is divided into three categories: The first category, less than thirty percent of people have hypertension or diabetes. In addition, they have no chance (a slim chance) of having hyperlipidemia. In the second group, more than 30 percent of people have hypertension or diabetes, and in the third group, more than 30 percent have hypertension or diabetes or hyperlipidemia. The accuracy of the experimental procedures is 80% (8 out of 10).

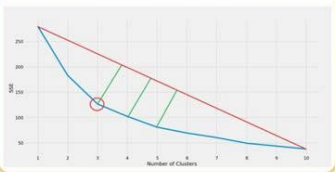
Flow Chart



Features

- Individual
  1. Age
  2. Sex
  3. Height
- BCS
  4. Weight
  5. Body Fat
  6. Muscle
  7. Visceral Fat

Elbow Method



1. Introduction

As most industrialized nations have already transitioned to an aging society, emerging nations are also aging. Mental, emotional, and physical decline are possible. Non-communicable diseases (NCDs) like diabetes, high blood pressure, hyperlipidemia, etc., should be monitored. It's the world's leading cause of death and disease burden. WHO expects NCD-related deaths to rise. Preventing and detecting symptoms early is crucial for timely care and reducing disease severity. This study's importance is clear. Using Body Composition Scale data to identify those at risk for prevalent NCDs by examining whether a factor is associated with the occurrence of a disease and putting it to the test using a data segmentation algorithm in the clustering of NCDs for surveillance, we can determine if a factor is relevant to the disease's occurrence. The results also determine future medical consultations.

2. Methodology

This independent study was conducted by collecting data from a sample of over-40s based on data from the InBody Dial Body Composition Analyzer, analyzing data from the sample group's disease survey, and observing trends of prevalent NCDs derived from **Fuzzy C-Means clustering** using the **Python Editor VS Code**.

2.1 Fuzzy C-Means Clustering

Minimize:

$$\sum_{j=1}^{k} \mu_{ij}^{p} \|x_i - c_j\|^2$$

Subject to:

$$\sum_{j=1}^{k} \mu_{ij} = 1$$

- x_i is an i-th dataset.
- c_j is a centroid of j-th cluster.
- μ_{ij} is a membership.
- k is a number of cluster.

2.2 Procedure

1. Randomly select μ_{ij} under conditions
2. Compute the centroids c_j using

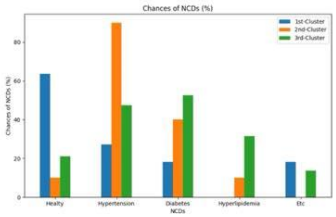
$$c_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{p} x_i}{\sum_{i=1}^{n} \mu_{ij}^{p}}$$

3. Calculate the membership μ_{ij} using

$$\mu_{ij} = \frac{\left(\frac{1}{\|x_i - c_j\|}\right)^{1/(p-1)}}{\sum_{j=1}^{k} \left(\frac{1}{\|x_i - c_j\|}\right)^{1/(p-1)}}$$

4. Repeat Step 2. and Step 3. until the membership μ_{ij} does not change.

3. Conclusions



Fuzzy C-Means clustering divides the data into three groups:

1. **less than 30% of people have hypertension or diabetes and have no chance (a slim chance) of having hyperlipidemia,**
2. **more than 30% of people have hypertension or diabetes,**
3. **more than 30% have hypertension or diabetes or hyperlipidemia.**

References

- Francios Chollet. "Deep Learning with Python", Second Edition, 2021
- Madson Luiz Dantas Dias, fuzzy c-means: An implementation of Fuzzy C-means clustering algorithm., Zenodo, 2019

Acknowledgement: HARIPHUNCHAI RAM HOSPITAL