การพัฒนาแนวทางการสรุปผลการตรวจสุขภาพประจำปี

DEVELOPMENT OF GUIDELINES FOR SUMMARIZING THE RESULTS OF ANNUAL HEALTH EXAMINATIONS

มงคล ตปนียปทุม

MONGKOL TAPANEEYAPATUM

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตร์มหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง
คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2566
KMITL-2023-SC-M-017-075

DEVELOPMENT OF GUIDELINES FOR SUMMARIZING THE RESULTS OF ANNUAL HEALTH EXAMINATIONS

MONGKOL TAPANEEYAPATUM

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE
IN DATA SCIENCE AND ANALYTICS
KMITL DIGITAL ANALYTICS AND INTELLIGENCE CENTER SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2023

KMITL-2023-SC-M-017-075

หัวข้อการค้นคว้าอิสระ การพัฒนาแนวทางการสรุปผลการตรวจสุขภาพประจำปี
ชื่อนักศึกษา นาย มงคล ตปนียปทุม
รหัสนักศึกษา 64605089
หลักสูตร วิทยาศาสตรมหาบัณฑิต (วิทยาการข้อมูลและการวิเคราะห์)
ศูนย์วิเคราะห์ข้อมูลอัจฉริยะพระจอมเกล้าลาดกระบัง
พ.ศ. 2566
อาจารย์ที่ปรึกษาการค้นคว้าอิสระ ผู้ช่วยศาสตราจารย์ ดร.ปัทมา เจริญพร

บทคัดย่อ

โครงงานค้นคว้าอิสระนี้นำเสนอแนวทางการสรุปผลการตรวจสุขภาพประจำปี โดยใช้ข้อมูลของโรงพยาบาลรามาธิบดีจักรีนฤบดินทร์ สถาบันการแพทย์จักรีนฤบดินทร์ โดยการนำองค์ความรู้จากการเรียนตลอดการศึกษามาประยุกต์ใช้

การพัฒนาโครงงาน มีจุดเริ่มต้นจาก นิสิตมีความสนใจในด้านความเสี่ยงในการเกิดโรคของพนักงานของโรงพยาบาลรามารธิบดี กับสถาบันการแพทย์จักรีนฤบดินทร์หลังจากที่พนักงานได้ทำการตรวจสุขภาพ เพื่อเป็นแนวทางการสรุปผลของการตรวจสุขภาพ ให้แพทย์ได้ให้คำปรึกษากับผู้ป่วยในการลด หรือควบคุมความเสี่ยงของผู้ป่วยในการเกิดโรคต่างๆ

หลังจากการตรวจสุขภาพประจำปี โครงงานนี้จึงเริ่มต้นพัฒนาขึ้นโดยการใช้ข้อมูลของผู้ป่วยที่มีอยู่ในปัจจุบัน โดยประยุกต์ใช้และพัฒนาความสามารถในการใช้องค์ความรู้ของนิสิต

ในการพัฒนาโครงงาน เริ่มตั้งแต่ การศึกษาและทำความใจข้อมูลของพนักงานภายในโรงพยาบาลรามาธิบดีจักรีนฤบดินทร์ ร่วมกับพนักของสถาบัน โดยแบ่งตามช่วงอายุ เพศ และดัชนีมวลกาย โดยการนำข้อมูลการตรวจสุขภาพประจำปี 2566 เพื่อนำมาเป็นข้อมูลขั้นต้นเพื่อให้ model เรียนรู้รูปแบบของข้อมูล เพื่อใช้ในการทำนาย โดยใช้ model ต่างๆ มาทำการทดลอง และทำนายโอกาสการเกิดโรคโรคต่างๆ ที่จะเกิดขึ้นกับพนักงานในอนาคต เพื่อให้แพทย์ให้คำแนะนำ เพื่อลดโอกาสในการเกิดโรคต่างๆ ในอนาคต และให้คำแนะนำเพื่อให้พนักงานมีสุขภาพดี

คำสำคัญ : ตรวจสุขภาพประจำปี

Independent Study Title	Development of Guidelines for Summarizing The Results of Annual Health Examinations
Student Name	Mongkol Tapaneeyapatum
Student ID	64605089
Degree	Master of Science (Data Science and Analytics) KMITL-Digital Analytics and Intelligence Center
Year	2023
Independent Study Advisor	Asst. Prof. Dr. Pattama Charoenporn

Abstract

This independent research project presents guidelines for summarizing the results of annual health examinations. Using data from Ramathibodi Chakri Naruebodindra Hospital Chakri Naruebodindra Medical Institute By applying knowledge from learning throughout the study.

project development Starting from Students are interested in the risk of disease among employees at Ramathibodi Hospital. with the Chakri Naruebodindra Medical Institute after the employee had a health check. To serve as a guideline for summarizing the results of the health examination Allow doctors to give advice to patients on reducing or control the patient's risk of developing various diseases

After the annual health examination This project was initially developed by using currently available patient data. By applying and developing students' ability to use knowledge.

In developing the project, it began with studying and understanding the information of employees within Ramathibodi Chakri Naruebodin Hospital. together with the staff of the institute Divided by age, gender, and body mass index. By using the 2023 annual health examination data to be used as primary data for the model to learn the patterns of the data. To be used in prediction by using various models to conduct experiments and predict the chance of developing various diseases. that will happen to employees in the future for the doctor to give advice To reduce the chance of developing various diseases in the future and provide advice to keep employees healthy.

Keywords : Annual Health Check Up

กิตติกรรมประกาศ

การค้นคว้าอิสระฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี ด้วยเพราะพระคุณของบิดามารดา นายบุญทิพย์ แซ่อึ้ง และ นางเส่ยม่วย แซ่อึ้ง ที่ให้กำลังใจ กำลังทรัพย์ และสนับสนุนทุกสิ่งทุกอย่างมาโดยตลอด ขอบคุณนางสาวกานต์กวิน สนองพัฒนกิจ และนางสาวกัญญา ตปนียปทุม น้องสาว ผู้สนันสนุน มอบกำลังใจ และความรักให้ผู้ศึกษา จนสามารถผ่านพ้นอุปสรรคต่างๆไปได้ด้วยใจที่แข็งแรง รวมถึงสมาชิกในครอบครัว ที่เป็นอีกกำลังใจและแรงสนับสนุนสำคัญตลอดการพัฒนาโครงงานค้นคว้าอิสระนี้

ขอขอบพระคุณท่านคณาจารย์ และวิทยากรทุกท่านที่มาถ่ายทอดความรู้ตลอดระยะเวลา 2 ปี ทำให้สามารถนำความรู้ต่างๆ ที่ได้เรียนมาใช้ประโยชน์ในการค้นคว้าอิสระนี้ได้อย่างเต็มที่ และความร่วมมือต่างๆ ของหลายท่าน ที่ให้การสนับสนุนข้าพเจ้า ตั้งแต่ต้นจนเสร็จสมบูรณ์ พร้อมทั้งขอบคุณเพื่อนๆ ทุกคนที่ให้ความช่วยเหลือทั้งในการเรียน งานกลุ่ม การสอบ และคำแนะนำ และช่วยหาข้อมูลที่จะนำมาใช้ในการค้นคว้าอิสระให้สำเร็จลุล่วงไปด้วยดี

ขอขอบพระคุณ ผศ. นพ.สิทธาคม ผู้สันติ, ผศ. นพ.จาตุรนต์ ตั้งสังวรธรรมะ, นพ.ชานน พุทธนวรัตน์ และพนักงานในงานสารสนเทศของสถาบันการแพทย์จักรีนฤบดินทร์ ที่ให้ความอนุเคราะห์แก่ข้าพเจ้าได้มีโอกาสศึกษาต่อในหลักสูตรนี้ ข้าพเจ้าหวังอย่างยิ่งที่จะนำความรู้ความสามารถที่มีต่อยอดและพัฒนางานให้เกิดประโยชน์แก่องค์กร

มงคล ตปนียปทุม

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

คำย่อ/สัญลักษณ์

คำย่อ/สัญลักษณ์	คำอธิบาย
LDL	Cholesterol Low-density Lipoprotein ไขมันไม่ดี
BMI	Body Mass Index ดัชนีมวลกาย
K-NN	K-Nearest Neighbour เพื่อนบ้านที่ใกล้เคียงที่สุด K ตัว
BS	ดาวสีน้ำเงิน
RC	วงกลมสีแดง
GS	สี่เหลี่ยมสีเขียว

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญ

ในปัจจุบันองค์กรณ์ต่างๆ ให้ความสำคัญกับสุขภาพของพนักงานมากขึ้น เพราะพนักงานเป็นปัจจัยสำคัญสำหรับทุกๆ องค์กรณ์ที่จะทำหน้าที่ในการพัฒนาองค์กรณ์ได้แบบยั่งยืน ดังนั้นการตรวจสุขภาพประจำปีจึงมีความสำคัญต่อองค์กรณ์ และพนักงาน เพราะเมื่อพนักงานมีสุขภาพแข็งแรงแล้วทำให้สามารถทำงานให้กับองค์กรณ์ได้อย่างมีประสิทธิภาพ

โครงการนี้เป็นการนำข้อมูลการตรวจสุขภาพประจำปีโดยมีช่วงอายุ เพศ และ BMI และค่า lab ต่างๆ เช่นค่า Cholesterol Low-density Lipoprotein (LDL) มาเพื่อนวิเคราะห์ข้อมูล โดยข้อมูลนี้นำมาจาก โรงพยาบาลรามาธิบดีจักรีนฤบดินทร์ สถาบันการแพทย์จักรีนฤบดินทร์ คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล เพื่อที่มาสรุปผลข้อมูลสุขภาพของพนักงานที่มาทำการตรวจในโรงพยาบาล

ปัจจุบัน ทางสถาบันการแพทย์จักรีนฤบดินทร์ ทำการรักษาโรคให้กับผู้ป่วยโดยดูจากประวัติการรักษาของผู้ป่วยเอง ไม่มีการเปรียบเทียบกับข้อมูลของผู้ป่วยอื่นๆ

1.2 วัตถุประสงค์ของงานวิจัย

1) เพื่อนำความรู้ด้านปัญญาประดิษฐ์มาประยุกต์ใช้ในการวิเคราะห์ และทำนายข้อมูลด้านสุขภาพของพนักงาน

2) เพื่อนำเสนอแบบจำลองที่สามารถใช้เป็นตัวช่วยในการแนะนำของแพทย์แก่พนักงานในด้านการรักษาสุขภาพ

3) เพื่อเปรียบเทียบการทำนายแบบเดิม จำลองแบบการทำนายแบบองค์รวมของแบบจำลอง

1.3 ขอบเขตของงานวิจัยที่ทำการศึกษา

งานวิจัยได้นำความรู้ด้านปัญญาประดิษฐ์มาประยุกต์ เพื่อพัฒนาแบบจำลองที่สามารถทำนายโอกาศการเกิดโรคที่เกี่ยวข้องกับการตรวจสุขภาพประจำปี

1) ข้อมูลที่นำมาใช้ในโครงการนี้ เป็นข้อมูลของในระบบสารสนเทศของ โรงพยาบาลรามาธิบดีจักรีนฤบดินทร์ สถาบันการแพทย์จักรีนฤบดินทร์ คณะแพทยศาสตร์ โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล

2) จำกัดสิทธิ์การเข้าถึงข้อมูล เฉพาะบุคลากรที่เกี่ยวข้อง หรือได้รับอนุญาติให้เข้าถึง จากสถาบันการแพทย์จักรีนฤบดินทร์

3) เป็นข้อมูลของพนักงานที่มาทำการตรวจสุขภาพประจำปีของ โรงพยาบาลรามาธิบดีจักรีนฤบดินทร์ และสถาบันการแพทย์จักรีนฤบดินทร์ ประจำปี 2566

1.4 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัยนี้

1) ได้นำความรู้ด้านปัญญาประดิษฐ์มาประยุกต์ใช้ในการวิเคราะห์ และทำนายข้อมูลด้านสุขภาพของการตรวจสุขภาพประจำปี

2) ได้เครื่องมือที่สามารถช่วยในการตัดสินใจในการวินิจฉัย ถึงโอกาสในการเกิดโรคต่างๆ ของพนักงาน เพื่อให้คำแนะนำกับพนักงานได้อย่างถูกต้อง

3) สามารถนำงานวิจัยมาต่อยอดกับผู้ป่วยที่อยู่ในองค์กรณ์อื่นๆ ว่าองค์กรณ์ควรรณรงค์เรื่องอะไรเพื่อให้สุขภาพของพนักงานในองค์กรณ์นั้น มีสุขภาพที่ดีพร้อมที่จะทำงาน เพื่อเป็นประโยชน์กับพนักงาน และองค์กรณ์

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง ซึ่งประกอบไปด้วย โรคอ้วน โรคที่เกี่ยวข้องกับไขมันสูง โรคไต โรคเบาหวาน ปัจจัยเสี่ยงของโรคสำคัญที่เกี่ยวข้องกับโรคไขมันสูง เบาหวาน เช่น โรคหัวใจและหลอดเลือด ความดันโลหิตสูง เป็นต้น ตัวชี้วัดทางเทคนิค แบบจำลองสำหรับการทำนายข้อมูล และงานวิจัยที่เกี่ยวข้อง ตามลำดับ

2.1 โรคอ้วน (Obesity)

ภาวะโรคอ้วนเกิดได้จากหลายปัจจัย โดยเฉพาะพฤติกรรมต่าง ๆ ของผู้คนที่เสี่ยงให้เกิดโรคอ้วน เช่น การเลือกรับประทานอาหารที่มีปริมาณคาร์โบไฮเดรตและน้ำตาลมากเกินไป ทั้งเบเกอรี่ ขนมหวาน ขนมกรุบกรอบ น้ำหวานชง น้ำอัดลม ฯลฯ ขาดการออกกำลังกายอย่างเหมาะสมเพียงพอ การลดความอ้วน หรือควบคุมน้ำหนักแบบไม่ถูกวิธีที่ทำให้กลับมาประสบปัญหาอ้วนมากกว่าเดิม บางรายอาจเกิดจากความผิดปกติของระบบเผาผลาญในร่างกายทำงานไม่สมบูรณ์ และอื่น ๆ

โรคอ้วนสามารถบ่งชี้ได้ด้วยการวัดค่าดัชนีมวลกาย คำนวณได้จากสมการ ดัชนีมวลกาย = น้ำหนัก (กิโลกรัม) / ส่วนสูง (เมตร^2)

$$BMI = \frac{Weight\ (kg)}{Height(m^2)}$$

ค่า BMI < 18.5	แสดงถึง	อยู่ในเกณฑ์น้ำหนักน้อยหรือผอม
ค่า BMI 18.5 – 22.90	แสดงถึง	อยู่ในเกณฑ์ปกติ
ค่า BMI 23 – 24.90	แสดงถึง	น้ำหนักเกิน
ค่า BMI 25 – 29.90	แสดงถึง	โรคอ้วนระดับที่ 1
ค่า BMI 30 ขึ้นไป	แสดงถึง	โรคอ้วนระดับที่ 2

โรคแทรกซ้อนต่าง ๆ ที่อาจมาพร้อมกับโรคอ้วน บางโรคอาจจะแสดงอาการให้เห็นชัดเจน ในขณะที่บางโรคก็อาจไม่แสดงอาการภายนอกแต่ส่งผลต่อสุขภาพในระยะยาว เช่นไขมันพอกตับ ภาวะหยุดหายใจขณะหลับ โรคหัวใจและหลอดเลือด โรคหลอเลือดสมอง กรดไหลย้อน ภาวะไอ จาม ปัสสาวะเล็ด ประจำเดือนผิดปกติ โรคเบาหวาน โรคข้อเสื่อม นอกจากนี้ยังมีโรคและภาวะอื่น ๆ เช่น ภาวะซึมเศร้าร่วมด้วย

2.2 โรคความดันโลหิตสูง (Hypertension)

เกิดจากการที่ความดันเลือดสูงกว่าปกติ คือความดันตัวบน (Systolic Blood Pressure) มากกว่า 140 และความดันตัวล่าง (Diastolic Blood Pressure) มากกว่า 90 ติดต่อกันเป็นเวลานาน มักไม่แสดงอาการ แต่ส่งผลเสียต่อหลอดเลือดและหัวใจ ทำให้เกิดภาวะแทรกซ้อนที่รุนแรงถึงชีวิต หรือพิการอาทิ โรคหัวใจ โรคหลอดเลือดสมอง กล้ามเนื้อหัวใจหนา เส้นเลือดแดงใหญ่โป่งพอง ไตวาย เป็นต้น การรู้ตัวว่าความดันโลหิตสูงตั้งแต่ระยะแรกนั้นสำคัญ ช่วยให้ควบคุมระดับความดันโลหิตและลดภาวะแทรกซ้อนที่รุนแรงได้

จากผลการสำรวจสุขภาพประชากรไทยโดยการตรวจร่างกายครั้งที่ 6 พ.ศ. 2562-2563 มีผู้ป่วยเป็นโรคความดันโลหิตสูง มากถึง 13 ล้านคน และกว่า 7 ล้านคนไม่ทราบว่าตนเองป่วย ที่มากรมควบคุมโลก กองโรคไม่ติดต่อ

ปัจจัยเสี่ยงความดันโลหิตสูง ความดันโลหิตสูงพบได้ 1 ใน 5 ของคนไทย จากการที่เส้นเลือดมีความเสื่อมตามวัย เมื่อความดันเพิ่มขึ้น เส้นเลือดจะแข็งและกระด้างมากขึ้น นอกจากนี้หากมีปัจจัยกระตุ้นอย่างกรรมพันธุ์ โรคเบาหวาน โรคอ้วน ภาวะไขมันในเลือดสูง การสูบบุหรี่ การดื่มแอลกอฮอล์ การกินอาหารรสจัด ความเครียด การพักผ่อนน้อย ยิ่งเสี่ยงความดันโลหิตสูงเพิ่มขึ้น

หากรู้เร็วและรักษาโรคความดันโลหิตสูงได้ไวจะช่วยลดการเกิดโรคหลอดเลือดสมองได้ 35 – 40% ลดการเกิดโรคหลอดเลือดหัวใจตัน 20 – 25% และลดการเกิดโรคหัวใจล้มเหลวได้มากกว่า 50% จึงควรวัดความดันอย่างสม่ำเสมอเพื่อตรวจเช็กร่างกายว่าความดันโลหิตอยู่ในเกณฑ์ปกติ

2.3 โรคไขมันในเลือดสูง (Dyslipidemia)

โรคที่มีระดับไขมันในเลือดสูงกว่าค่าที่ถูกกำหนดขึ้น ซึ่งได้มาโดยการเก็บข้อมูลทางสถิติของระดับไขมันในเลือดของประชากรทั่วไป ปกติร่างกายคนเราจะมีไขมันอยู่ 2 ชนิด คือ

2.3.1 คอเลสเตอรอล (Cholesterol)

แบ่งเป็น ชนิดความหนาแน่นต่ำ (LDL) หรือ ไขมันชนิดไม่ดี เป็นคอเลสเตอรอลที่ไปสะสมในผนังหลอดเลือด ทำให้หลอดเลือดแดงตีบและแข็ง เป็นสาเหตุของโรคหลอดเลือดหัวใจตีบตัน และหลอดเลือดสมองตีบ

ชนิดความหนาแน่นสูง (HDL) หรือ ไขมันชนิดดี เป็นคอเลสเตอรอลประเภทหนึ่งเหมือนกัน แต่จะทำหน้าที่กำจัดไขมันชนิดอันตรายออกไปจากกระแสเลือด ซึ่งช่วยลดความเสี่ยงในการเกิดโรคเส้นเลือดหัวใจตีบ

2.3.2 ไตรกลีเซอไรด์ (Triglyceride)

เป็นไขมันประเภทหนึ่ง ซึ่งอาจมีการสะสมที่ผนังหลอดเลือดได้เช่นกันเมื่อมีปริมาณสูงมาก ๆ แต่จะมีอิทธิพลต่อการเกิดโรคหัวใจและหลอดเลือดน้อยกว่าไขมันชนิดคอเลสเตอรอล เกณฑ์ปกติ ระดับไขมันในร่างกายคือ

โคเลสเตอรอลรวม ต่ำกว่า 200 mg/dL

ไตรกลีเซอไรด์ ต่ำกว่า 150 mg/dL

ไขมันชนิดไม่ดี LDL ต่ำกว่า 130 mg/dL

ในผู้ชาย ไขมันชนิดดี HDL สูงกว่า 40 mg/dL

ในผู้หญิง ไขมันชนิดดี HDL สูงกว่า 50 mg/dL

2.4 โรคไต (Kidney disease)

เกิดจากการที่ไตทำงานผิดปกติ สามารถเกิดขึ้นได้ทุกเพศ ทุกวัย โดยเฉพาะผู้สูงอายุ แม้ว่าหนึ่งในสาเหตุของการเกิดโรคไตนั้นมาจากการกินเค็ม แต่การไม่กินเค็ม ไม่เติมเกลือ หรือน้ำปลาในอาหาร ก็ไม่ได้หมายความว่าคุณจะไม่เสี่ยงเป็นโรคไต เนื่องจากโรคไตนั้นไม่ได้เกิดจากพฤติกรรมการกินอาหารที่มีรสเค็มเพียงอย่างเดียวเท่านั้น แต่ยังมีสาเหตุอื่นๆ ทั้งจากการใช้ชีวิตประจำวัน พันธุกรรม และโรคเรื้อรัง ก็ทำให้มีโอกาสป่วยเป็นโรคไตได้เช่นกัน

สำหรับในประเทศไทย มีผู้ป่วยไตเรื้อรัง จำนวน 11.6 ล้านคน และมีจำนวนมากกว่า 1 แสนคนที่ต้องล้างไต และจากรายงานของ The United States Renal Data System (USRDS) พบว่าประเทศไทยเป็น 1 ใน 5 ประเทศที่มีอัตราการเกิดโรคไตสูงที่สุด ที่มา กรมอนามัย / 9 มีนาคม 2566

ระยะของโรคไตเรื้อรัง

โรคไตเรื้อรังแบ่งเป็น 5 ระยะ โดยแบ่งตามระดับของอัตราการกรองของไต (eGFR)

ระยะของโรคไตเรื้อรัง	eGFR (mL/min/1.73 m)	คำนิยาม
ระยะที่ 1	> 90	มีภาวะไตผิดปกติ เช่น มีโปรตีนรั่วในปัสสาวะ แต่อัตราการกรองยังปกติ
ระยะที่ 2	60 - 90	มีภาวะไตผิดปกติ เช่น มีโปรตีนรั่วในปัสสาวะ อัตราการกรองลดลงเล็กน้อย
ระยะที่ 3a	45 - 59	อัตราการกรองลดลงเล็กน้อยถึงปานกลาง
ระยะที่ 3b	30 - 44	อัตราการกรองลดลงปานกลางถึงมาก
ระยะที่ 4	15 - 29	อัตราการกรองลดลงมาก
ระยะที่ 5	< 15	ไตวายระยะสุดท้าย

ค่า eGFR (estimated Glomerular Filtration Rate) คือปริมาณเลือดที่ไหลผ่านตัวกรองของไตในหนึ่งนาที (มล./นาที/1.73 ตร.ม.)

รูปที่ 2.1 โรคไตเรื้อรังแบ่งเป็น 5 ระยะโดยแบ่งตามระดับของอัตราการกรองของไต

สาเหตุของการเป็นโรคไตไม่ใช่แค่กินเค็มเท่านั้น แต่ยังมีปัจจัยอื่นๆ ที่มีส่วนสำคัญในการกระตุ้นให้ร่างกายคนเรามีโอกาสเสี่ยงที่จะเกิดโรคไต ได้เหมือนกัน โดยปัจจัยที่พบบ่อยได้แก่

การมีโรคที่มีผลกระทบกับไต เช่น เบาหวาน ความดันโลหิตสูง โรคหัวใจและหลอดเลือด โรคอ้วนหรือมีภาวะน้ำหนักเกิน โรคเก๊าท์หรือระดับกรดยูริกในเลือดสูง โรคแพ้ภูมิตนเอง การสูบบุหรี่เรื้อรัง ซึ่งจะส่งผลให้หลอดเลือดที่ไปเลี้ยงไตเสื่อมลง ส่งผลให้การทำงานของไตเสื่อมลง

การมีภาวะไตผิดปกติ ไม่สมบูรณ์ตั้งแต่กำเนิด เช่น ไตฝ่อ มีมวลเนื้อไตลดลง หรือมีไตข้างเดียว เป็นต้น

การมีภาวะหลอดเลือดฝอยในไตอักเสบ

การเป็นโรคติดเชื้อทางเดินระบบปัสสาวะส่วนบนซ้ำหลายครั้ง

การตรวจพบนิ่วในไตหรือระบบทางเดินปัสสาวะ หรือตรวจพบถุงน้ำในไตมากกว่า 3 ตำแหน่งขึ้นไป

การมีประวัติครอบครัวเป็นโรคไตเรื้อรัง หรือ มีประวัติการเป็นโรคไตอักเสบ หรือถุงน้ำในไต

การได้รับยาในกลุ่มยาแก้ปวดกลุ่ม NSAIDs หรือสารพิษที่ทำลายไต (Nephrotoxic agents)

ดื่มน้ำน้อยเกินไป เกิดภาวะขาดน้ำของไตจนทำงานบกพร่อง หรือเกิดการสะสมของสารเคมีในทางเดินปัสสาวะ จนตกตะกอนกลายเป็นโรคนิ่วในไตหรือทางเดินปัสสาวะ

รับประทานอาหารที่มีโซเดียมสูงแต่ไม่เค็ม อาทิ ซอสมะเขือเทศ ซอสพริก น้ำจิ้มสุกี้ อาหารแปรรูปยอดนิยม เช่น แฮม เบคอน ขนมกรุบกรอบ ผลไม้กระป๋อง รวมถึงอาหารหมักดอง เช่น ผักกาดดอง หรือไข่เค็ม บะหมี่กึ่งสำเร็จรูป ผงฟู สารกันบูด หรือสารกันเชื้อราในขนมปัง เป็นต้น

2.5 โรคเบาหวาน

โรคเบาหวานเป็นความผิดปกติของระบบเผาผลาญชนิดหนึ่ง เป็นเพราะการขับอินซูลินภายในไม่เพียงพอหรือทำงานผิดปกติทำให้เกิดความผิดปกติของการเผาผลาญของสารอาหาร เช่น คาร์โบไฮเดรต โปรตีน และไขมัน ทำให้มีกลูโคสในเลือดมากเกินไป ซึ่งขับออกจากร่างกายด้วยปัสสาวะทางไต ส่งผลให้มีน้ำตาลในปัสสาวะ

การวินิจฉัยโรคเบาหวานขึ้นอยู่กับความหนาแน่นของกลูโคสในเลือดเป็นหลัก โดยการตรวจหลังจากการอดอาหาร 8 ชั่วโมง ระดับน้ำตาลในเลือดของผู้ใหญ่ปกติคือ 70–110 มก./ดล. และระดับน้ำตาลในเลือดภายใน 2 ชั่วโมงหลังอาหารจะน้อยกว่า 140 ก./ดล. (Chimei, 2007) อัตราการเสียชีวิตของผู้ป่วยเบาหวานที่ขึ้นกับอินซูลินอยู่ที่ประมาณ 6 เท่าของเพศชายในวัยเดียวกัน และ 10 เท่าของเพศหญิงในวัยเดียวกัน (Science News, 1990)

โรคอ้วนเกิดจากการสะสมของไขมันในอวัยวะภายในมากเกินไป ดังนั้นจึงมีความสัมพันธ์เชิงบวกกับโรคเบาหวานและหลอดเลือด นอกจากนี้ ค่าดัชนีมวลกาย WHR และรอบเอวสูงมีความสัมพันธ์อย่างมากกับอัตราการเกิดโรคเบาหวาน และผู้ที่มีดัชนีมวลกายสูงและ WHR มีความเสี่ยงต่อโรคเบาหวาน (Faster, 1983; Pouliot et al., 1994)

2.6 โรคหัวใจและหลอดเลือด

โรคหัวใจหมายถึงการไหลเวียนของเลือดไม่เพียงพอเนื่องจากความผิดปกติของเนื้อเยื่อหลอดเลือดหรือการบดเคี้ยว และหัวใจไม่สามารถรับออกซิเจนเพียงพอ ดังนั้นกล้ามเนื้อหัวใจบางส่วนจะขาดออกซิเจนหรือตาย ส่งผลต่อการทำงานทั่วไป อาการทางคลินิกเนื่องจากกล้ามเนื้อหัวใจมีไม่เพียงพอ ได้แก่ ภาวะหัวใจเต้นผิดจังหวะ โรคหลอดเลือดหัวใจตีบ การอุดตันของหัวใจและหลอดเลือด ภาวะหัวใจล้มเหลว และการเสียชีวิตกะทันหันที่ไม่มีอาการ อาการทางคลินิก ได้แก่ อาการเจ็บหน้าอก แรงกดที่หน้าอกด้านหน้าซ้าย ทรวงอก หายใจลำบากหรือรู้สึกไม่ย่อย ใจสั่น เหงื่อออกเย็น มีอาการวิงเวียนศีรษะซับซ้อน อาการอ่อนเปลี้ยเพลียแรง

ปัจจัยเสี่ยงที่นำไปสู่โรคหัวใจและหลอดเลือด มีปัจจัยเสี่ยงมากกว่า 300 ชนิด ที่สัมพันธ์ต่อการเกิดโรคหัวใจและหลอดเลือด อย่างไรก็ตามองค์การอนามัยโลกระบุว่า ผู้ที่ป่วยด้วยโรคนี้อย่างน้อยหนึ่งในสาม เกี่ยวข้องกับการสูบบุหรี่ เบาหวาน ความดันเลือดสูง คอเลสเตอรอลสูง และความอ้วน ซึ่งเป็นสิ่งที่พบได้ในชีวิตของคนทั่วไป ด้วยเหตุนี้คนทั่วไปที่มีปัจจัยเสี่ยงเหล่านี้จึงมีโอกาสเป็นโรคนี้ได้ ไม่ว่าจะเป็นคนฐานะดีหรือไม่ก็ตาม

ปัจจัยเสี่ยงเหล่านี้เกี่ยวข้องกับคนทุกเพศ ทุกวัย และเป็นสิ่งที่สามารถหลีกเลี่ยงปรับเปลี่ยนและควบคุมได้ทั้งสิ้น อย่างไรก็ตาม จากสถานการณ์สุขภาพของคนไทยในปัจจุบันกำลังปรากฏแนวโน้มปัจจัยเสี่ยงสำคัญเหล่านี้เพิ่มขึ้น โดยเฉพาะอย่างยิ่งในกลุ่มเด็กและวัยรุ่น

จากการที่โรคหัวใจและหลอดเลือดเป็นผลจากความสัมพันธ์ระหว่างปัจจัยเสี่ยงหลายปัจจัยประกอบเข้าด้วยกัน การป้องกันและรักษาให้ได้ผลจึงไม่อาจมุ่งไปที่ปัจจัยอย่างหนึ่งอย่างใด แต่ต้องพิจารณร่วมกันทั้งหมด และต้องป้องกันอย่างต่อเนื่องตั้งแต่วัยเด็กไปจนตลอดชีวิตจึงได้ผลดีที่สุด

การกินผักและผลไม้ให้เพียงพอ ออกกำลังกาย หรือมีกิจกรรมทางกายที่มากพอ และไม่สูบบุหรี่ช่วยลดความเสี่ยงต่อการเกิดโรคหลอดเลือดหัวใจตีบตันจนกล้ามเนื้อหัวใจตาย (myocardial infarction) ได้ถึงร้อยละ 80

2.7 โรคหลอดเลือดสมอง

โรคหลอดเลือดสมองเป็นชนิดของการเปลี่ยนแปลงทางพยาธิวิทยาในหลอดเลือดสมอง และภาวะแทรกซ้อนของเส้นโลหิตตีบหลอดเลือดแดงทั่วไป เส้นโลหิตตีบในสมองนำไปสู่หลอดเลือดตีบหรือการลอกของคราบพลัคหลอดเลือดโดยไม่ได้ตั้งใจ และปิดกั้นหลอดเลือดสมองที่อยู่ห่างไกลออกไป และอาจนำไปสู่เส้นเลือดอุดตันในสมอง กล้ามเนื้อหัวใจตาย หรือหลอดเลือดสมองแตกและมีเลือดออก ในความเป็นจริง ผู้ป่วยโรคหลอดเลือดสมองมักจะมีโรคเรื้อรังอื่นๆ ในระดับต่างๆ เช่น ความดันโลหิตสูง หลอดเลือดหัวใจตีบ ไขมันในเลือดสูง กรดยูริกในเลือดสูง เบาหวาน และโรคอ้วน นอกจากนี้ โรคหลอดเลือดสมองและโรคหัวใจและหลอดเลือดมีปฏิกิริยาต่อกันทั้งสาเหตุ และผลการวิจัยทางคลินิกเรื่อง "REACH Registry" ได้ตรวจสอบผู้ป่วยนอกที่มีความเสี่ยงสูง 67,800 รายใน 44 ประเทศต่อปี พบว่า 40% ของผู้ป่วยโรคหลอดเลือดสมองมีเส้นเลือดอุดตันที่หลอดเลือดหัวใจหรือหลอดเลือดส่วนปลาย 25% ของผู้ป่วยหลอดเลือดหัวใจมีเส้นเลือดอุดตันในสมองหรือหลอดเลือดแดงส่วนปลาย (Carmen, 2007) นอกจากนี้ หากผู้ป่วยมีโรคเกี่ยวกับหัวใจ เช่น โรคหลอดเลือดหัวใจตีบ ภาวะหัวใจเต้นผิดจังหวะ หรือโรคลิ้นหัวใจ ภาวะดังกล่าวก็จะเกี่ยวเนื่องกันได้ง่ายจากโรคหลอดเลือดสมอง เช่น เส้นเลือดอุดตันหรือเส้นเลือดสมองตีบ

ปัจจัยเสี่ยงของโรคหลอดเลือดสมองสามารถแบ่งออกได้เป็นปัจจัยเสี่ยงหลัก ๆ ได้แก่ ผู้สูงอายุ ความดันโลหิตสูง โรคหัวใจ เบาหวาน ภาวะชักจากสมองขาดเลือดชั่วขณะและประวัติโรคหลอดเลือดสมอง และปัจจัยเสี่ยงรอง ได้แก่ ไขมันในเลือดสูง โรคอ้วน ภาวะเม็ดเลือดแดงมากผิดปกติ การสูบบุหรี่ การดื่มสุรา , กรรมพันธุ์ในครอบครัว ยาคุมกำเนิด และยาอื่นๆ

จากทฤษฎีการเกิดโรคต่างๆ ที่เกี่ยวข้องกับการตรวจสุขภาพประจำปี จะเห็นว่าในการเกิดโรคชนิดหนึ่ง อาจจะเป็นปัจจัยเสี่ยงทำให้เกิดโรคต่างๆ ขึ้นมาอีกหลายโรค ดังนั้นจึงควรหมั่นรักษาสุขภาพให้ดีอยู่เสมอ เช่นการรับประทานอาหารให้ครบ 5 หมู่ และไม่มากเกินไปจนทำให้เกิดภาวะโรคอ้วนขึ้น ออกกำลังกาย คลายความเครียด พักผ่อนให้เพียงพออยู่เสมอ ก็จำทำให้เรารักษาสุขภาพที่ดีได้

2.8 เทคโนโลยีการทำเหมืองข้อมูล

เทคโนโลยีการจำแนกประเภทการทำเหมืองข้อมูลประกอบด้วยสองส่วน: การสร้างแบบจำลองการจำแนกประเภท และการประเมินประสิทธิภาพการจำแนกแบบจำลอง ในส่วนแรกการจำแนกประเภทที่นำมาใช้อัลกอริธึมได้รับการฝึกอบรมโดยชุดข้อมูลการฝึกอบรมที่เป็นความลับ เพื่อสร้างแบบจำลองการทำนายการจำแนกประเภท ในส่วนที่สอง ชุดข้อมูลการทดสอบใช้เพื่อทดสอบประสิทธิภาพการจำแนกประเภทของโมเดลนี้ ทุกข้อมูลในชุดข้อมูลการฝึกอบรมหรือชุดข้อมูลการทดสอบมีจำนวนแอตทริบิวต์และคลาสเป้าหมายต่างกัน การศึกษานี้ใช้การตรวจสอบความถูกต้องแบบ 10-fold cross validation ในการสร้างแบบจำลองการจำแนกประเภทและการประเมินประสิทธิภาพ

และอัลกอริธึมการจำแนกประเภทใช้โครงสร้างการตัดสินใจ ตัวแยกประเภทแบบเบย์ และโครงข่ายประสาทแบบกระจายกลับ

2.8.1. อัลกอริทึมแผนผังการตัดสินใจ

โครงสร้างการตัดสินใจเป็นชนิดของการจำแนกและคาดการณ์เทคโนโลยีการทำเหมืองข้อมูล ซึ่งเป็นของการเรียนรู้แบบอุปนัยและเทคโนโลยีการขุดความรู้ภายใต้การดูแล เนื่องจากแผนผังการตัดสินใจมีประโยชน์ในการสร้างอย่างรวดเร็วและสร้างกฎการตัดสินใจแบบ if-then ที่ตีความได้ง่าย จึงกลายเป็นเทคนิคที่ใช้กันอย่างแพร่หลายมากที่สุดในบรรดาวิธีการจำแนกประเภทต่างๆ (Cabena et al., 1997; Kennedy, Lee, Roy, Reed, & ลิปมัน, 1997)

โครงสร้างการตัดสินใจเป็นชนิดของแผนภาพต้นไม้ โหนดที่ด้านบนของโครงสร้างต้นไม้คือโหนดราก โหนดด้านล่างคือโหนดปลายสุด และโหนดระดับเป้าหมายหนึ่งรายการจะถูกกำหนดให้กับโหนดปลายสุดแต่ละโหนด

ตั้งแต่โหนดรูทไปจนถึงโหนดปลายสุดทุกโหนด มีพาธที่สร้างจากโหนดภายในหลายโหนดพร้อมแอตทริบิวต์ เส้นทางนี้สร้างกฎที่จำเป็นสำหรับการจัดประเภทข้อมูลที่ไม่รู้จัก นอกจากนี้ อัลกอริธึมทรีการตัดสินใจส่วนใหญ่ยังมีงานสองขั้นตอน เช่น การสร้างต้นไม้และการตัดแต่งต้นไม้

ในขั้นตอนการสร้างต้นไม้ อัลกอริธึมทรีการตัดสินใจสามารถใช้วิธีการ (ฟังก์ชัน) ที่เป็นเอกลักษณ์ในการเลือกแอตทริบิวต์ที่ดีที่สุด เพื่อแยกชุดข้อมูลการฝึกอบรม สถานการณ์สุดท้ายของขั้นตอนนี้คือข้อมูลที่อยู่ในชุดย่อยการฝึกแยกเป็นของคลาสเป้าหมายเพียงคลาสเดียวเท่านั้น การเรียกซ้ำและการทำซ้ำตามการเลือกแอตทริบิวต์และการแยกการตั้งค่าจะเติมเต็มการสร้างโหนดรากของโครงสร้างการตัดสินใจและโหนดภายใน ในทางกลับกัน ข้อมูลพิเศษบางอย่างในชุดข้อมูลการฝึกอบรมอาจนำไปสู่สาขาที่ไม่เหมาะสมบนโครงสร้างแผนผังการตัดสินใจ ซึ่งเรียกว่าการใส่มากเกินไป ดังนั้น หลังจากสร้างแผนผังการตัดสินใจแล้ว จะต้องตัดแต่งกิ่งเพื่อขจัดกิ่งที่ไม่เหมาะสม เพื่อเพิ่มความแม่นยำของแบบจำลองการตัดสินใจในการทำนายข้อมูลใหม่ (Quinlan, 1986; Witten & Frank, 2000)

ในบรรดาอัลกอริธึมแผนผังการตัดสินใจที่พัฒนาแล้ว อัลกอริทึมที่ใช้กันทั่วไป ได้แก่ ID3 (Maher & Clair, 1993), C4.5 (Breiman, Friedman, Olshen, & Stone, 1984), CART (Kass, 1980) และ CHAID (Quinlan, 1993) C4.5 ได้รับการพัฒนาจากอัลกอริธึม ID3 (Iterative Dichotomiser 3) โดยใช้ทฤษฎีข้อมูลและวิธีการเรียนรู้แบบอุปนัยเพื่อสร้างแผนผังการตัดสินใจ C4.5 ปรับปรุง ID3 ซึ่งไม่สามารถประมวลผลปัญหาตัวเลขต่อเนื่องได้ อัลกอริธึม CHAID มีจุดเด่นในการใช้การทดสอบไคสแควร์เพื่อคำนวณค่า p ของหมวดหมู่โหนดในการ

แยกทุกครั้ง เพื่อกำหนดว่าจะอนุญาตให้แผนภูมิการตัดสินใจเติบโตโดยไม่ต้องตัดแต่งหรือไม่ CHAID ไม่สามารถประมวลผลข้อมูลต่อเนื่องได้ ดังนั้นจึงใช้ไม่ได้กับปัญหาทางการแพทย์จำนวนมากที่มีข้อมูลตัวเลขต่อเนื่อง อัลกอริธึม CART เป็นวิธีการแยกไบนารี ใช้ในข้อมูลที่แอตทริบิวต์ต่อเนื่องกัน ดัชนี Gini ใช้เพื่อประเมินดุลยพินิจของข้อมูลเป็นพื้นฐานในการเลือกเงื่อนไขการแยก เนื่องจากการศึกษาครั้งนี้เป็นการประมวลผลข้อมูลทางการแพทย์ที่มีคุณลักษณะหลายอย่าง จึงเลือก C4.5 เป็นอัลกอริทึมแผนผังการตัดสินใจ

อัลกอริธึมทรีการตัดสินใจถูกนำมาใช้ในงานทางการแพทย์หลายอย่าง เช่น ในการเพิ่มคุณภาพของการวินิจฉัยโรคผิวหนัง (Chang & Chen, 2009), การทำนายความดันโลหิตสูงที่จำเป็น (Ture, Kurt, Kurum, & Ozdamar, 2005) และการทำนายโรคหลอดเลือดหัวใจ (ออม คิม และจาง 2008)

2.8.2 การแยกประเภทแบบเบย์

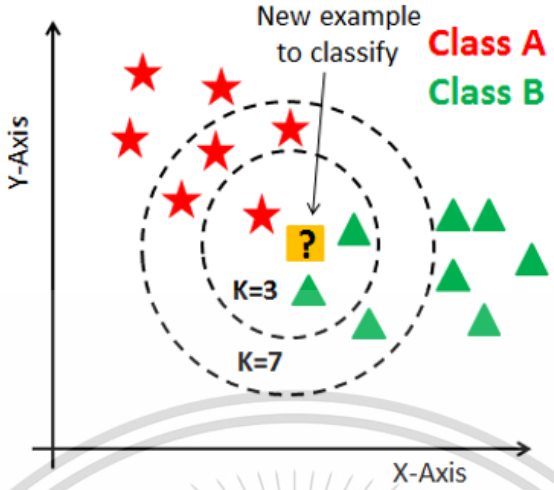
ทฤษฎีการแยกประเภทแบบเบย์เกิดจากทฤษฎีบทแบบเบย์ในสถิติ ในขณะที่ตั้งสมมติฐานไว้ล่วงหน้า กล่าวคือ ทุกแอตทริบิวต์มีความเป็นอิสระ เพื่อให้ตัวแยกประเภทสามารถทำได้ง่ายและรวดเร็ว ตามทฤษฎีบทเบย์ ความน่าจะเป็นของชุดข้อมูล X_t ที่เป็นของ C คือ: $P(C|X_t) = \frac{p(C)p(X_t|C)}{p(X_t)}$

ตามสูตรข้างต้น ตัวแยกประเภทแบบเบย์จะคำนวณความน่าจะเป็นแบบมีเงื่อนไขของอินสแตนซ์ที่เป็นของแต่ละคลาส และตามข้อมูลความน่าจะเป็นแบบมีเงื่อนไขดังกล่าว อินสแตนซ์จะถูกจัดประเภทเป็นคลาสที่มีความน่าจะเป็นแบบมีเงื่อนไขสูงสุด ในการแสดงออกของความรู้ มีการตีความที่ยอดเยี่ยมเช่นเดียวกับแผนผังการตัดสินใจ และสามารถใช้ข้อมูลก่อนหน้าเพื่อสร้างแบบจำลองการวิเคราะห์สำหรับการทำนายหรือการจัดประเภทในอนาคต (Loether & McTavish, 1993) หากค่าลักษณะเฉพาะของข้อมูลมีความต่อเนื่อง มีสองวิธีในการประมวลผล (Vapnik, 1982)

2.8.2.1 สมมติว่าเป็นการแจกแจงแบบปกติและหา (หมายถึง ความแปรปรวน) ของค่าลักษณะเฉพาะเป็นความน่าจะเป็น

2.8.2.2 ใช้วิธีแยกเพื่อถ่ายโอนข้อมูลต่อเนื่องเป็นข้อมูลที่ไม่ต่อเนื่อง

การแยกประเภทแบบเบย์ถูกนำมาใช้ในประเด็นทางการแพทย์มากมาย เช่น การวัดคุณภาพการดูแลผู้ป่วยฉุกเฉินทางจิตเวช เป็นต้น (Gustafson, Sainfort, Johnson, & Sateia, 1993) ช่วยในการวินิจฉัย มะเร็งเต้านม (Wang, Zheng, Good, King, & Chang, 1999) และกรณีทางการแพทย์ (Kononenko, 1993).



รูปที่ 2.2 เพื่อนบ้านใกล้ที่สุด k ตัว (KNN)

2.8.3 เพื่อนบ้านที่ใกล้ที่สุด k ตัว (KNN)

ทฤษฎีการหา K-Nearest Neighbor (KNN) เป็นเทคนิคการเรียนรู้ของเครื่องยอดนิยมที่ใช้สำหรับงานจำแนกประเภทและการถดถอย ขึ้นอยู่กับแนวคิดที่ว่าจุดข้อมูลที่คล้ายคลึงกันมักจะมีป้ายกำกับหรือค่าที่คล้ายคลึงกัน

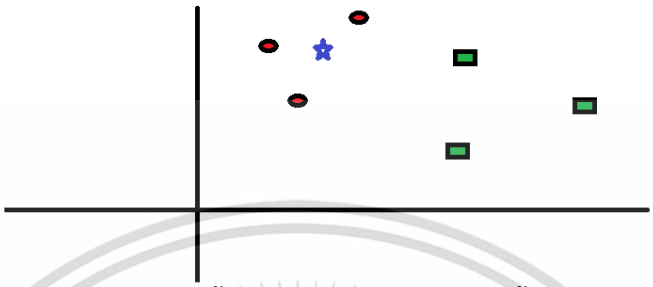
ในระหว่างขั้นตอนการฝึก อัลกอริธึม KNN จะจัดเก็บชุดข้อมูลการฝึกทั้งหมดไว้เป็นข้อมูลอ้างอิง เมื่อทำการพยากรณ์ ระบบจะคำนวณระยะห่างระหว่างจุดข้อมูลอินพุตและตัวอย่างการฝึกทั้งหมด โดยใช้การวัดระยะทางที่เลือก เช่น การวัดระยะทางด้วย Euclidean Distance หากเรามีข้อมูลจำนวนมาก การนำข้อมูลไปสร้างกราฟแล้วใช้การกะด้วยสายตาในการแบ่งข้อมูล ก็จะให้คำตอบที่ตอบที่ไม่แม่นยำนัก เราจึงจะใช้การหาระยะทางที่เรียกว่า Euclidean Distance ในการช่วยหาว่าข้อมูลที่เราต้องการใกล้กับข้อมูลประเภทไหนมากกว่ากัน เมื่อเราต้องการหาระยะทางจากจุด (x1,y1) ไป (x2,y2) โดยกำหนดให้ d แทนระยะทาง จะได้ว่า $d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$

ดังนั้น $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

ถัดไป อัลกอริธึมจะระบุ K เพื่อนบ้านที่ใกล้ที่สุดกับจุดข้อมูลอินพุตตามระยะทาง ในกรณีของการจำแนกประเภท อัลกอริธึมจะกำหนดป้ายกำกับคลาสที่พบมากที่สุดในหมู่เพื่อนบ้าน K เป็นป้ายกำกับที่คาดการณ์สำหรับจุดข้อมูลอินพุต สำหรับการถดถอย จะคำนวณ

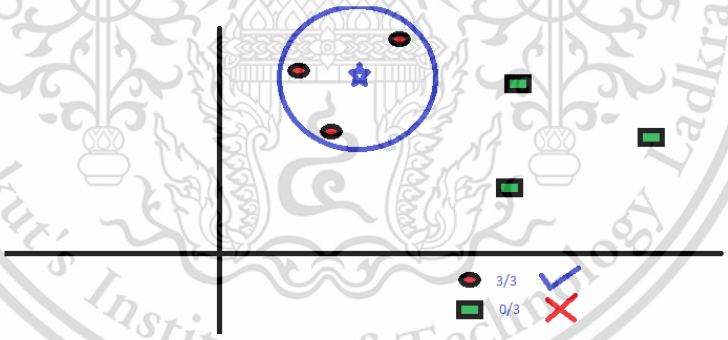
ค่าเฉลี่ยหรือค่าเฉลี่ยถ่วงน้ำหนักของค่าเป้าหมายของเพื่อนบ้าน K เพื่อทำนายค่าสำหรับจุดข้อมูลอินพุต

อัลกอริทึม KNN ทำงานอย่างไร ลองใช้กรณีง่ายๆ เพื่อทำความเข้าใจอัลกอริทึมนี้ ต่อไปนี้เป็นการแพร่กระจายของวงกลมสีแดง (RC) และสี่เหลี่ยมสีเขียว (GS)



รูปที่ 2.3 กำหนด class ดาวสีน้ำเงิน ในกราฟเพื่อเป็นตัวตั้งต้นหา k เพื่อนบ้านใกล้ที่สุด

คุณตั้งใจที่จะค้นหาคลาสของดาวสีน้ำเงิน (BS) BS อาจเป็น RC หรือ GS ก็ได้ และไม่มีอะไรอื่นอีก อัลกอริธึม “K” ใน KNN คือเพื่อนบ้านที่ใกล้ที่สุดที่เราประสงค์จะลงคะแนน สมมติว่า K = 3 ดังนั้น ตอนนี้เราจะสร้างวงกลมโดยให้ BS เป็นจุดศูนย์กลางที่ใหญ่พอๆ กับการล้อมจุดข้อมูลเพียงสามจุดบนระนาบ โปรดดูแผนภาพต่อไปนี้สำหรับรายละเอียดเพิ่มเติม



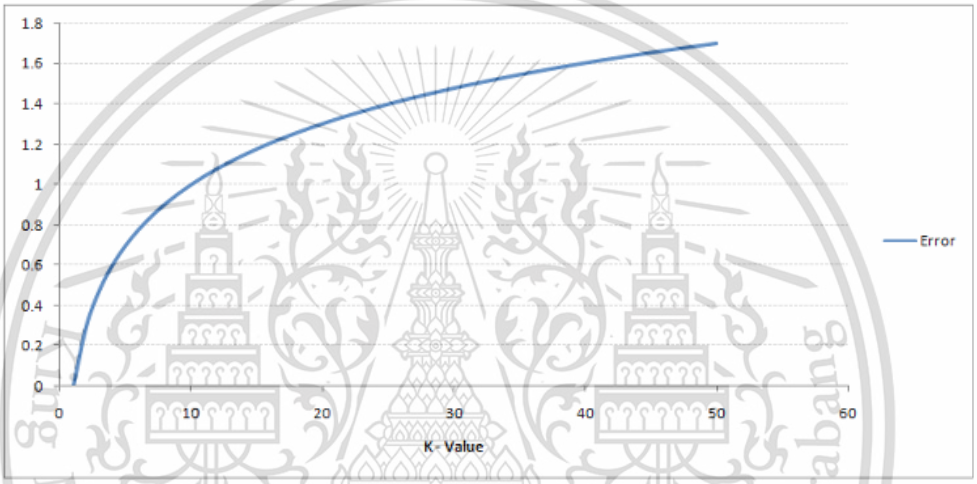
รูปที่ 2.4 หาค่าที่ใกล้ดาวสีน้ำเงิน มากที่สุด โดยกำหนดให้ k = 3

สามจุดที่ใกล้เคียงที่สุดกับ BS คือ RC ทั้งหมด ดังนั้น ด้วยระดับความมั่นใจที่ดี เราสามารถพูดได้ว่า BS ควรอยู่ในคลาส RC ที่นี่ ตัวเลือกปรากฏชัดเจนเมื่อทั้งสามโหวตจากเพื่อนบ้านที่ใกล้ที่สุดไปที่ RC การเลือกพารามิเตอร์ K มีความสำคัญมากในอัลกอริทึมนี้ ต่อไปเราจะมาทำความเข้าใจปัจจัยที่ต้องพิจารณาเพื่อสรุป K ที่ดีที่สุด

เราจะเลือกปัจจัย K ได้อย่างไร ก่อนอื่นให้เราพยายามทำความเข้าใจอย่างชัดเจนถึงอิทธิพลของ K ในอัลกอริธึม หากเราเห็นตัวอย่างสุดท้าย เมื่อพิจารณาว่าการสังเกตการฝึกทั้ง 6 ครั้งยังคงที่ ด้วยค่า K ที่กำหนด เราสามารถสร้างขอบเขตของแต่ละคลาสได้ ขอบเขตการ

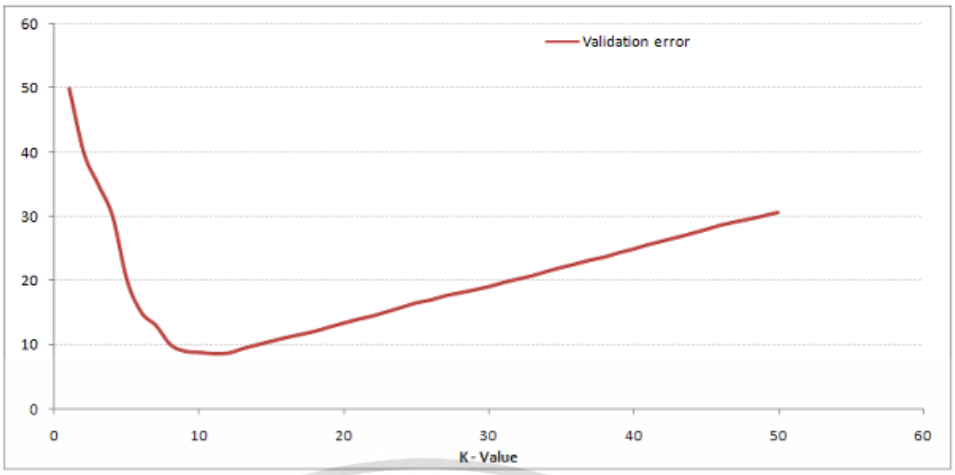
ตัดสินใจเหล่านี้จะแยก RC ออกจาก GS ในทำนองเดียวกัน เรามาดูผลกระทบของค่า “K” ต่อขอบเขตของคลาสกัน ต่อไปนี้เป็นขอบเขตที่แตกต่างกันซึ่งแยกทั้งสองคลาสด้วยค่า K ที่ต่างกัน

หากคุณสังเกตดีๆ คุณจะเห็นว่าขอบเขตจะราบรื่นขึ้นเมื่อค่า K เพิ่มขึ้น เมื่อ K เพิ่มขึ้นจนถึงระยะอนันต์ ในที่สุดมันจะกลายเป็นสีน้ำเงินทั้งหมดหรือสีแดงทั้งหมดขึ้นอยู่กับเสียงข้างมากทั้งหมด อัตราความผิดพลาดในการฝึกและอัตราความผิดพลาดในการตรวจสอบความถูกต้องเป็นพารามิเตอร์สองตัวที่เราต้องใช้เพื่อเข้าถึงค่า K ที่แตกต่างกัน ต่อไปนี้คือเส้นโค้งสำหรับอัตราความผิดพลาดในการฝึกที่มีค่าแปรผันเป็น K



รูปที่ 2.5 การเพิ่มค่า k และอัตราการเพิ่มขึ้นของ Error

อย่างที่คุณเห็น อัตราข้อผิดพลาดที่ K=1 จะเป็นศูนย์เสมอสำหรับตัวอย่างการฝึก เนื่องจากจุดที่ใกล้เคียงที่สุดกับจุดข้อมูลการฝึกคือตัวมันเอง ดังนั้นการทำนายจึงแม่นยำเสมอด้วย K=1 หากเส้นโค้งข้อผิดพลาดในการตรวจสอบความถูกต้องจะคล้ายกัน ตัวเลือก K ของเราจะเป็น 1 ต่อไปนี้คือเส้นโค้งข้อผิดพลาดในการตรวจสอบความถูกต้องที่มีค่าต่างกันของ K



รูปที่ 2.6 กราฟการวัดความถูกต้องของ error

ทำให้เรื่องราวมีความชัดเจนมากขึ้น ที่ K=1 เรากำลังทำเกินขอบเขต ดังนั้นอัตราข้อผิดพลาดเริ่มแรกจะลดลงและถึงระดับต่ำสุด หลังจากจุดต่ำสุด จากนั้นจะเพิ่มขึ้นตามค่า K ที่เพิ่มขึ้น เพื่อให้ได้ค่า K ที่เหมาะสมที่สุด คุณสามารถแยกการฝึกอบรมและการตรวจสอบความถูกต้องออกจากชุดข้อมูลเริ่มต้นได้ ตอนนี้พล็อตเส้นโค้งข้อผิดพลาดในการตรวจสอบความถูกต้องเพื่อให้ได้ค่าที่เหมาะสมที่สุดของ K ค่า K นี้ควรใช้สำหรับการทำนายทั้งหมด

2.8.4. วิธีการตรวจสอบแบบ 10-fold cross validation

ตามอัตราส่วนข้อมูลหมวดหมู่ดั้งเดิม แบ่งชุดข้อมูลทดลองแบบสุ่มออกเป็น 10 ชุดย่อยข้อมูลที่เท่ากัน ใช้ชุดย่อยข้อมูล 9 ชุดสำหรับชุดข้อมูลการฝึกอบรม และชุดย่อยข้อมูลที่เหลือเป็นชุดข้อมูลการทดสอบ ทำซ้ำ 10 ครั้ง อนุญาตให้ทุกชุดย่อยของข้อมูลทำหน้าที่เป็นข้อมูลการทดสอบ และใช้ค่าเฉลี่ยของผลการทดสอบ 10 รายการเพื่อประเมินประสิทธิภาพของแบบจำลองการคาดการณ์

2.9 การทำ features selection

2.9.1 การเลือกแบบ Information Gain

เป็นการประเมินค่าเพื่อใช้ในการแบ่งข้อมูลด้วยการ คำนวณค่า Gain สาหรับแต่ละมิติข้อมูลถ้ามิติข้อมูลใดมีค่า Gain สูงสุด จะถูกเลือกให้เป็น features

2.9.2 การเลือกแบบ LASSO

จะเป็นการทำ Regularization ที่แตกต่างจาก Ridge regression เพราะว่า penalty term นั้นไม่ได้นำค่า coefficient มายกกำลัง ข้อดีของ Lasso คือ จะมีการกดตัวแปรที่ไม่สำคัญ และเหลือเฉพาะตัวแปรที่เด่นๆ เท่านั้น

$$\beta(\lambda) = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|^2 + \lambda|\beta|$$

รูปที่ 2.7 สูตรทั่วไปของ model LASSO

2.9.3 การเลือกแบบรีเคอร์ซิฟฟิทเจอร์อิลิมิเนชัน (Recursive Feature Elimination)

หลักการทำงานในการเลือกฟีเจอร์ ก็คือ เริ่มต้นด้วยการใช้ฟีเจอร์ทั้งหมดก่อน และทำการกำหนดว่าต้องการ feature กี่ตัว จากนั้นทำการคำนวณค่าความสำคัญของฟีเจอร์แต่ละตัว โดยใชฟังชันก์ของ sklearn เช่น coef_ หรือ feature_importance_ และทำการ ตัดฟีเจอร์ที่มีความเกี่ยวข้องน้อยที่สุดออกไป เพื่อทำให้ขนาดของฟีเจอร์ลดลง จนกว่าจะได้มาซึ่งชุดของฟีเจอร์ที่เล็กที่สุด แต่ยังคงมีประสิทธิภาพการทำงานที่ดีนั่นเอง ลักษณะการทำงานเป็นแบบวนซ้ำ (recursive) หลาย ๆ รอบ

2.10 งานวิจัยที่เกี่ยวข้องกับระบบการทำนายโอกาสการเกิดโรค หรือมีปัญหาสุขภาพ

ในบทย่อยนี้จะกล่าวถึงงานวิจัยที่เกี่ยวข้องกับระบบการทำนายโอกาสการเกิดโรคหลอดเลือดสมอง เพื่อนำองค์ความรู้มาประยุกต์ใช้กับโครงงาน ซึ่งผู้พัฒนาได้เลือกหัวข้อวิจัยที่สนใจ ทั้งหมด 3 โครงงาน ได้แก่ A predictive model for cerebrovascular disease using data mining, Incidence and risk factors for stroke in patients with COVID-19 in the Philippines: An analysis of 10,881 cases, Cardiovascular risk factors and 10-year CV risk scores in adults aged 30-70 years old in Amnat Charoen Province, Thailand

2.10.1 A predictive model for cerebrovascular disease using data mining

โรคหลอดเลือดสมองจัดอยู่ในอันดับที่ 2 หรือ 3 ของสาเหตุการเสียชีวิต 10 อันดับแรกในไต้หวัน และคร่าชีวิตผู้คนไปประมาณ 13,000 รายทุกปีตั้งแต่ปี 1986 เมื่อโรคหลอดเลือดสมองเกิดขึ้น ไม่เพียงแต่นำไปสู่ค่ารักษาพยาบาลที่สูงเท่านั้น แต่ยังถึงขั้นเสียชีวิตด้วย

ประเทศที่พัฒนาแล้วทั้งหมดในโลกให้ความสำคัญกับการป้องกันและรักษาโรคหลอดเลือดสมอง ทุ่มงบประมาณและทรัพยากรมนุษย์จำนวนมากในการศึกษาระยะยาว เพื่อลดภาระหนัก เนื่องจากการเกิดโรคของโรคหลอดเลือดสมองมีความซับซ้อนและผันแปร การวินิจฉัยที่แม่นยำล่วงหน้าจึงเป็นเรื่องยาก อย่างไรก็ตาม ในมุมมองของเวชศาสตร์ป้องกันจำเป็นต้องสร้างแบบจำลองการทำนายเพื่อปรับปรุงการวินิจฉัยโรคหลอดเลือดสมองที่แม่นยำ ดังนั้น ร่วมกับโปรแกรมการป้องกันและรักษาโรคหลอดเลือดสมองปี 2550 ของโรงพยาบาลเพื่อการสอนระดับภูมิภาคในไต้หวัน การศึกษาครั้งนี้จึงมุ่งที่จะใช้เทคโนโลยีการจำแนกประเภทเพื่อสร้างแบบจำลองการทำนายโรคหลอดเลือดสมองที่เหมาะสมที่สุด จากแบบจำลองการทำนายนี้ ได้ดึงกฎการจำแนกโรคหลอดเลือดสมองมาใช้เพื่อปรับปรุงการวินิจฉัยและการทำนายโรคหลอดเลือดสมอง

2.10.2 Incidence and risk factors for stroke in patients with COVID-19 in the Philippines: An analysis of 10,881 cases

ในขณะที่การศึกษาขนาดใหญ่ส่วนใหญ่เกี่ยวกับความสัมพันธ์ที่เป็นไปได้ของ COVID-19 และโรคหลอดเลือดสมองนั้นดำเนินการในประเทศที่มีรายได้สูง แต่มีการศึกษาเพียงไม่กี่ชิ้นที่ประกอบด้วยกลุ่มตัวอย่างขนาดเล็กที่ทำขึ้นในประเทศที่มีรายได้ต่ำถึงปานกลาง เช่น ฟิลิปปินส์

เพื่อกำหนดปัจจัยเสี่ยงของโรคหลอดเลือดสมองในผู้ป่วย COVID-19 ที่รักษาในโรงพยาบาลในฟิลิปปินส์ เพื่อกำหนดความสัมพันธ์ที่เป็นไปได้ระหว่างปัจจัยเสี่ยงเหล่านี้กับโรคหลอดเลือดสมองในกลุ่มประชากรเดียวกัน และเพื่อพิจารณาว่ามีความเกี่ยวข้องกันระหว่างการตายกับโรคหลอดเลือดสมองในกลุ่มเดียวกันนี้หรือไม่

2.10.3 Cardiovascular risk factors and 10-year CV risk scores in adults aged 30-70 years old in Amnat Charoen Province, Thailand

โรคหัวใจและหลอดเลือด (CVD) เป็นสาเหตุการเสียชีวิตอันดับหนึ่งของโลก กว่าสามในสี่ของเหตุการณ์ CVD เกิดขึ้นในประเทศที่มีรายได้น้อยและปานกลาง และอัตรา CVD ในประเทศไทยนั้นสูงกว่าเมื่อเทียบกับประเทศอื่นๆ ในเอเชียที่ 32.3% งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาคะแนนความเสี่ยงโรคหัวใจและหลอดเลือดของคนไทยในประชากรอายุ 30-70 ปี ที่อาศัยอยู่ในจังหวัดอำนาจเจริญ จังหวัดภาคตะวันออกเฉียงเหนือของประเทศไทย การศึกษาแบบภาคตัดขวางนี้ใช้การสุ่มตัวอย่างที่ดำเนินการโดยความน่าจะเป็นตามสัดส่วนกับขนาด โดยมีผู้เข้าร่วม 382 คนจาก 2 อำเภอจาก 7 อำเภอของอำนาจเจริญ เก็บรวบรวมข้อมูลโดยใช้การวัดสัดส่วนร่างกาย การตรวจเลือด และการสัมภาษณ์แบบตัวต่อตัว ที่การตรวจวัดพื้นฐานคะแนนความเสี่ยง CV เฉลี่ยของไทยอยู่ในระดับต่ำ (6.1 ± SD 5.5%) การวิจัยแสดงให้เห็นว่า

83.0 เปอร์เซ็นต์ของผู้เข้าร่วมมีคะแนนความเสี่ยง CV ต่ำ ในขณะที่ 17.0 เปอร์เซ็นต์มีคะแนนความเสี่ยง CV สูง การวิเคราะห์การถดถอยโลจิสติกแสดงให้เห็นว่าตัวทำนายที่สำคัญของคะแนนความเสี่ยง CV คืออายุ (Odds Ratio (OR) 81.74; 95% Confidence Interval (CI) 28.65-233.18) ความดันโลหิตซิสโตลิก (SBP) (OR, 5.90; 95% CI 2.28- 15.29), สถานะการสูบบุหรี่ (OR, 4.12; 95% CI 1.15-14.69), เวลาเฉลี่ยในการนอนตอนกลางคืน (OR, 2.81; 95% CI 1.17-7.09) และการมีส่วนร่วมในกิจกรรมด้านสุขภาพ (OR, 3.89; 95% CI 1.62-9.36). ผลการวิจัยนี้บ่งชี้ว่าพฤติกรรมบางอย่างอาจนำไปสู่เหตุการณ์ CVD ดังนั้นจึงแนะนำให้นำผลการศึกษาเหล่านี้ไปปรับใช้ในการสร้างแนวทางการเน้นย้ำการใช้ชีวิตอย่างมีสุขภาพในการบำบัดรักษา และกำหนดมาตรการที่เหมาะสมหรือโปรแกรมสุขศึกษาเพื่อสร้างความรู้และจิตสำนึกเกี่ยวกับเหตุการณ์ CVD ของผู้ใหญ่ชาวไทย และเนื่องด้วย 24% ของผู้ป่วยที่เป็นโรคหลอดเลือดหัวใจ มีโอกาสที่จะเป็นโรคหลอดเลือดสมองตีบจึงมีความสนใจในงานวิจัยนี้

บทที่ 3
วิธีการดำเนินงานวิจัย

ในงานวิจัยนี้ผู้วิจัยได้แบ่งขั้นตอนในการดำเนินการวิจัยออกเป็น 4 ขั้นตอนประกอบด้วย 1) การเตรียมข้อมูล และการกำหนดคำนิยามให้กับข้อมูล 2) การพัฒนาเครื่องมือสำหรับจำแนกประเภทผู้ป่วย 3) การพัฒนาเครื่องมือสำหรับการจำแนกประเภทแนวโน้มสุขภาพโดยรวมของพนักงาน 4) การวัดประสิทธิภาพแบบจำลองในการในจำแนกประเภทสุขภาพของผู้ป่วย และ 5) เครื่องมือที่ใช้สำหรับการวิจัย

3.1 การเตรียมข้อมูล และการกำหนดคำนิยามให้กับข้อมูล

3.1.1 การเตรียมข้อมูล

ในขั้นตอนนี้ จะนำข้อมูลของผู้ป่วยที่มารับการตรวจเช็คสุขภาพประจำปี 2566 โดยนำมาจากระบบสารสนเทศของโรงพยาบาลรามาธิบดีจักรีนฤบดินทร์ โดยเป็นค่า lab และกำหนดประเภทของข้อมูล

Class	ผลของ Class
1	ปกติ
2	ไขมันผิดปกติ
3	ไตผิดปกติ
4	น้ำตาลผิดปกติ
5	ระดับฮีโมโกลบินผิดปกติ
6	ความดันผิดปกติ
7	BMI ผิดปกติ
8	มีความผิดปกติมากกว่า 2 ค่า

รูปที่ 3.1 การแปลผลของ class ที่กำหนดในข้อมูล

โดยมีข้อมูลทั้งหมดที่มาจากฐานข้อมูลระบบสารสนเทศของ โรงพยาบาลรามาธิบดีจักรีนฤบดินทร์ 1,285 row 88 columns แต่ทางผู้วิจัย เลือกมาใช้งาน 14 columns

```
1 df1 = df[cols]
2 print(df1.shape)
```

```
(1285, 14)
```

```
1 df1.head(3)
```

	class	gender_group	age_group	bmi_val	sbp	dbp	hb_val	fbs_val	creatinine_val	egfr_val	triglyceride_val	t_cholesterol_val	hdl_val	ldl_val
0	1	2	1	24.0	113	74	13.1	0	0.61	126.9	105	170	49	110
1	8	1	1	27.0	112	75	15.2	0	0.95	106.2	109	208	38	161
2	8	2	1	19.0	105	69	11.9	0	0.49	136.4	82	170	44	114

รูปที่ 3.2 จำนวนข้อมูลที่เลือกมาใช้ในการทำวิจัย

3.1.2 การหา Outlier

จากนั้นทำ remove outliers โดยใช้ค่า bmi_val upper bound = Q3 + (4*IQR) lower bound = Q1 – (1*IQR) จะทำการลบข้อมูลที่มีค่า bmi_val มากกว่า 54 และมีค่าน้อยกว่า 12

```
1 #detect and remove outliers
2 Q3, Q1 = np.percentile(df1['bmi_val'],[75,25])
3 IQR = Q3 - Q1
4 upper_bound = Q3 + (4*IQR)
5 lower_bound = Q1 - (1*IQR)
6
7 df1 = df1.loc[(df1['bmi_val']>=lower_bound)&(df1['bmi_val']<=upper_bound)]
8 df2 = df2.loc[(df2['bmi_val']>=lower_bound)&(df2['bmi_val']<=upper_bound)]
```

```
1 print('Upper bound : = ',upper_bound)
2 print('Lower bound : = ',lower_bound)
3 print('df Shape : = ',df.shape)
4 print('df1 Shape : = ',df1.shape)
5 print('df2 Shape : = ',df2.shape)
```

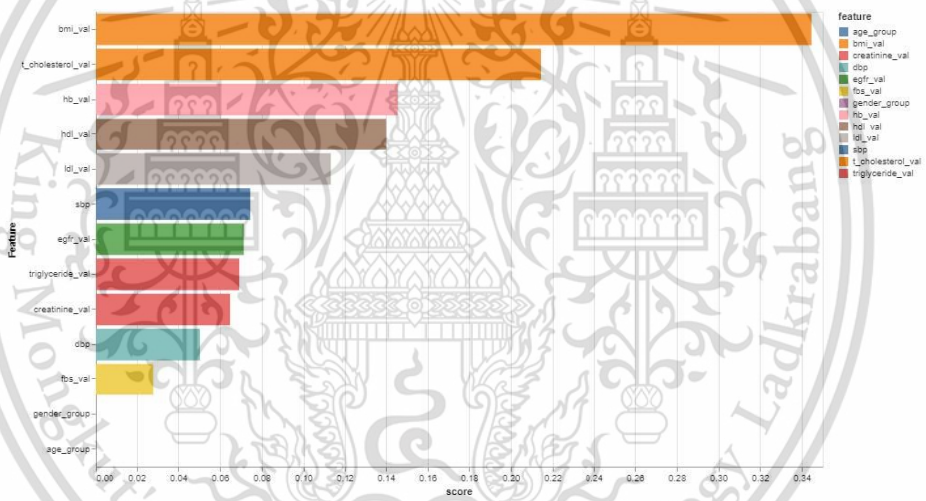
```
Upper bound : =  54.0
Lower bound : =  12.0
df Shape : =  (1285, 88)
df1 Shape : =  (1158, 14)
df2 Shape : =  (1158, 14)
```

รูปที่ 3.3 สูตรการคำนวณหาค่า outlier

หลังจากเรา remove outlier แล้วจะทำการนำข้อมูลที่เหลือไปใช้งาน โดยการหา outlier = upper bound + lower bound = 127 แถวข้อมูลก่อนหา outlier = 1285 – outlier = 127 เหลือข้อมูลที่จะนำมาใช้งาน = 1158 แถว จากนั้นเรานำข้อมูลมาหา feature selection โดยใช้ Information Gain ใช้ model = mutual_info_regression

ตารางที่ 3.1 แสดงค่า Information Gain โดยใช้ model = mutual_info_regression

index	Feature	score
0	bmi_val	0.344584
1	t_cholesterol_val	0.214394
2	hb_val	0.145758
3	hdl_val	0.140163
4	ldl_val	0.113133
5	Sbp	0.074519
6	egfr_val	0.071393
7	triglyceride_val	0.069238
8	creatinine_val	0.064821
9	Dbp	0.050333
10	fbs_val	0.027916
11	gender_group	0
12	age_group	0



รูปที่ 3.4 กราฟแสดงค่า Information Gain โดยใช้ model = mutual_info_regression

3.1.3 การทำ features selection

จากนั้นเราลองมา feature selection โดยการใช้ Lasso Cross Validation, Recursive Feature Eliminator and Random Forest Regressor เรานำมา combining 3 model เพื่อหา feature selection โดยการ votes ในการใช้ Lasso Cross Validation เราสามารถหามาได้ 10 features ['bmi_val', 'sbp', 'hb_val', 'fbs_val', 'creatinine_val', 'egfr_val', 'triglyceride_val', 't_cholesterol_val', 'hdl_val', 'ldl_val'], เมื่อใช้ Recursive Feature Eliminator and Random Forest Regressor เรากำหนดไว้ 10 features จะได้ผลตามรูปที่ 3.5 จะเหลือ 8 features

Combining 3 feature selectors

```
# Sum the votes of the three models
votes1 = np.sum([lcv_mask1, rf_mask1, gb_mask1], axis=0)
print(votes1)

# Create a mask for features selected by all 3 models
meta_mask1 = votes1 == 3
print(meta_mask1)

# Apply the dimensionality reduction on X
X1_reduced = X1_train.loc[:, meta_mask1]
print(X1_reduced.columns)
```

```
[0 0 3 3 2 3 3 2 3 2 3 3 3]
[False False  True  True False  True  True False  True False  True  True
  True]
Index(['bmi_val', 'sbp', 'hb_val', 'fbs_val', 'egfr_val', 't_cholesterol_val',
       'hdl_val', 'ldl_val'],
      dtype='object')
```

รูปที่ 3.5 ผล feature selection โดยการ combinig 3 model

3.2 ทำการจำลองโดยให้ model ทำการเรียนรู้

เราทำการจำลอง model โดยใช้ 3 model Naïve Bayes, K-Nearest Neighbors [KNN] and Tensorflow Decision Forests

ตารางที่ 3.2 เปรียบเทียบค่า accuracy ระหว่าง 3 model

Model \ Features selections	None = 13 Features	Combine = 8 Features
Naïve Bayes	0.6293	0.6954
K-Nearest Neighbors [KNN]	0.5373	0.5431
Decision Tree	0.9321	0.8938

จากการทดลอง ปรากฏว่า model Decision Tree มีค่าความถูกต้องมากที่สุด โดยการที่ไม่ลด Features ที่เราเลือกมาใช้ในการทดลองผ่าน model

ตารางต่อจากนี้เป็นตารางข้อมูลของการวัดประสิทธิภาพของ model 3 model โดยเปรียบเทียบเป็น confusion matrix

	actual positive	actual negative
predicted positive	TP	FP
predicted negative	FN	TN

รูปที่ 3.6 ตาราง confusion matrix

Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$

Recall = $\frac{TP}{TP+FN}$

Precision = $\frac{TP}{TP+FP}$

F1-score = $\frac{2\ x\ Precision\ x\ Recall}{Precision+\ Recall}$

ตารางที่ 3.3 ตัววัดประสิทธิภาพของ model Naïve Bayes ก่อนทำ features selection

	Actual								
	1	2	3	4	5	6	7	8	Precision
1	12	5	3	0	0	0	11	1	0.48
2	6	68	2	0	1	0	2	20	0.59
3	1	0	4	0	0	0	1	0	0.4
4	1	0	1	0	0	0	0	0	0
5	1	0	0	0	3	0	3	1	0.5
6	0	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	8	4	0.2
8	4	43	0	0	2	0	15	124	0.82
Recall	0.38	0.69	0.67	0	0.38	0	0.67	0.66	

ตารางที่ 3.4 ตัววัดประสิทธิภาพของ model Naïve Bayes หลังทำ features selection

	Actual								
	1	2	3	4	5	6	7	8	Precision
1	15	5	2	0	1	1	6	2	0.48
2	6	71	0	0	0	0	1	21	0.67
3	3	0	3	0	0	0	0	0	0.60
4	1	0	0	0	0	0	0	1	0.00
5	2	0	0	0	4	0	2	0	0.57
6	0	0	0	0	0	0	0	1	0.00
7	3	0	0	0	0	0	6	3	0.22
8	1	30	0	0	2	0	2	143	0.84
Recall	0.53	0.39	0.17	0	0	1	0.17	0.68	

ตารางที่ 3.5 ตัววัดประสิทธิภาพของ model [KNN] ก่อนทำ features select

	Actual								
	1	2	3	4	5	6	7	8	Precision
1	17	3	1	0	0	0	8	3	0.37
2	5	39	0	0	0	0	1	54	0.45
3	3	1	1	0	0	0	1	0	0.25
4	2	0	0	0	0	0	0	0	0
5	5	2	0	0	0	0	1	0	0
6	0	0	0	0	0	1	0	0	1
7	5	1	1	0	1	0	2	2	0.1
8	9	41	1	0	3	0	7	127	0.68
Recall	0.53	0.39	0.17	0	0	1	0.17	0.68	

ตารางที่ 3.6 ตัววัดประสิทธิภาพของ model [KNN] หลังทำ features select

	Actual								
	1	2	3	4	5	6	7	8	Precision
1	21	1	0	0	3	1	5	1	0.44
2	2	50	0	0	0	0	1	46	0.48
3	4	0	0	0	0	0	1	1	0.00
4	1	0	0	0	0	0	1	0	0.00
5	4	2	0	0	1	0	1	0	0.12
6	0	0	0	0	0	1	0	0	0.33
7	6	2	0	0	0	0	3	1	0.14
8	10	50	0	0	4	1	10	113	0.70
Recall	0.66	0.51	0.00	0.00	0.12	1.00	0.25	0.60	

ตารางที่ 3.7 ตัววัดประสิทธิภาพของ model Decision Tree ก่อนทำ features select

	Actual								
	1	2	3	4	5	6	7	8	Precision
1	30	2	0	0	0	0	0	0	0.83
2	0	96	0	0	0	0	0	3	0.94
3	4	0	2	0	0	0	0	0	1.00
4	2	0	0	0	0	0	0	0	0.00
5	0	0	0	0	6	0	0	2	1.00
6	0	0	0	0	0	0	0	1	0.00
7	0	0	0	0	0	0	8	4	0.73
8	0	4	0	0	0	0	3	181	0.95
Recall	0.94	0.97	0.33	0.00	0.75	0.00	0.67	0.96	

ตารางที่ 3.8 ตัววัดประสิทธิภาพของ model Decision Tree หลังทำ features select

	Actual								
	1	2	3	4	5	6	7	8	Precision
1	30	2	0	0	0	0	0	0	0.44
2	1	97	0	0	0	0	0	1	0.48
3	3	0	3	0	0	0	0	0	0.00
4	2	0	0	0	0	0	0	0	0.00
5	0	0	0	0	5	0	0	3	0.12
6	0	0	0	0	0	0	0	1	0.33
7	0	0	0	0	0	0	8	4	0.14
8	0	9	0	0	2	0	3	174	0.70
Recall	0.94	0.98	0.50	0.00	0.62	0.00	0.67	0.93	

3.3 ทำการวัดประสิทธิภาพ

ในการวัดประสิทธิภาพเพื่อทำการเปรียบเทียบแต่ละ model เราจะใช้ค่า accuracy, precision และ recall ว่า model ไหนควรนำมาใช้งาน ตามตารางที่ 3.3 – 3.8

บทที่ 4
ผลการวิจัยและการอภิปรายผล

วัตถุประสงค์หลักของงานวิจัยนี้คือการนำความรู้ทางด้านปัญญาประดิษฐ์ โดยเฉพาะเทคนิคการเรียนรู้ของเครื่อง และการเรียนรู้เชิงลึก ได้แก่ แบบจำลอง Naïve Bayes, แบบจำลองเพื่อนบ้านใกล้ที่สุด k ตัว และแบบจำลองต้นไม้ตัดสินใจ เพื่อพัฒนาเป็นแบบจำลองสำหรับทำนายการเกิดโรค ที่สัมพันธ์กับการตรวจร่างกาย ซึ่งแบบจำลองนี้จะช่วยให้ผู้ได้รับการตรวจ แพทย์ สามารถกำหนดทิศทางการการปฏิบัติตัวของผู้ที่มารับการตรวจสุขภาพได้ นอกจากนี้ยังได้นำเสนอวิธีการเพิ่มประสิทธิภาพให้กับแบบจำลองด้วยเทคนิคต่างๆ ซึ่งจะทำให้แบบจำลองทำนายข้อมูลได้อย่างแม่นยำมากขึ้น นอกจากการวัดประสิทธิภาพด้านความแม่นยำในการทำนายแบบจำลองแล้ว งานวิจัยนี้ได้ทำการวัดประสิทธิภาพในด้านการทำกำไรของแบบจำลองต่างๆ จากผลการทดลองทั้งหมด สามารถสรุปผลการศึกษาและวิเคราะห์ผลการศึกษาที่ได้ดังต่อไปนี้

4.1 ผลการเตรียมข้อมูล

4.1.1 การเตรียมข้อมูลพื้นฐาน

ผู้วิจัยเตรียมข้อมูลพื้นฐานโดยใช้ Library pandas ฟังก์ชัน read_csv() เพื่อทำการอ่านข้อมูลในไฟล์ CSV ที่ได้ import มาจากโปรแกรม HIS ซึ่งได้แก่ข้อมูลของพนักงานสถาบันจักรีนฤบดินทร์ และโรงพยาบาลรามาธิบดีที่มาทำการตรวจสุขภาพประจำปี พ.ศ. 2566 โดยมีรายละเอียดดังรูปที่ 4.1

```
1 # Load raw data in csv file
2 df = pd.read_csv("/content/drive/MyDrive/IS_dataset/checkup_raw_data_10.csv",engine="python",encoding = "UTF-8")
```

```
1 df.head(3)
```

	id	[redacted]	class	class_label	gender_desc	gender_group	age	group_of_age	...	U_WBC_rec	U_Squamous_epithelial_cell_val	U_Squamous_epithelial_cell_rec	U_Bacteria_val	U_Bacteria_rec	U_Mucous_Thread
0	65172		1	ปกติ	Female	2	24	น้อยกว่า 35 ปี	...	ปกติ	0-1	ปกติ	Few	ไม่แปลผล	Neg
1	64881		8	มีความผิดปกติมากกว่า 1 ค่า	Male	1	31	น้อยกว่า 35 ปี	...	ปกติ	Negative	ปกติ	Negative	อยู่ในเกณฑ์ปกติ	Neg
2	64887		8	มีความผิดปกติมากกว่า 1 ค่า	Female	2	24	น้อยกว่า 35 ปี	...	ผิดปกติ	03-05	ปกติ	Moderate	ไม่แปลผล	Neg

3 rows × 88 columns

รูปที่ 4.1 ตัวอย่างข้อมูลที่นำมาใช้ในการเรียนรู้

ก่อนที่จะนำเข้าข้อมูลเพื่อให้แบบจำลองเรียนรู้ต้องมีการทำความสะอาดข้อมูล (Data Cleaning) จะอยู่ในขั้นตอน Data preprocessing

```
1 df_class_label.value_counts()
```

```
class  class_label
8      มีความผิดปกติมากกว่า 1 ค่า    758
2      ไขมันผิดปกติ                  295
1      ปกติ                           96
7      BMI ผิดปกติ                    85
5      ระดับฮีโมโกลบินผิดปกติ         28
3      ไตผิดปกติ                      17
6      ความดันผิดปกติ                 4
4      น้ำตาลผิดปกติ                  2
dtype: int64
```

รูปที่ 4.2 จำนวน class ที่มีการกำหนดใน dataset

จากนั้นจึงเลือก features จากdataset ทั้งหมด โดยกำหนดไว้ 13 columns มีดังนี้

'class', = การแปลผลการตรวจสุขภาพ

'gender_group', = เพศของผู้รับการตรวจสุขภาพ

'age_group', = ช่วงอายุของผู้รับการตรวจสุขภาพ (1 = อายุน้อยกว่า 35 ปี, 2 = อายุ 35-39 ปี, 3 = อายุ 40-49 ปี, 4 = อายุ 50-59 ปี, 5 = อายุ 60 และ 60 ปีขึ้นไป)

'bmi_val', = ช่วงของ bmi จะแบ่งเป็น 5 ระดับ (1 = อยู่ในเกณฑ์น้ำหนักน้อยหรือผอม, 2 = อยู่ในเกณฑ์ปกติ, 3 = น้ำหนักเกิน, 4 = โรคอ้วนระดับที่ 1, 5 = โรคอ้วนระดับที่ 2)

ตารางที่ 4.1 การคำนวณค่า bmi พร้อมแปลค่า

ค่า BMI = นน.(kg.)/ส่วนสูง(m^2)	หมายความว่า
ค่า BMI < 18.5	อยู่ในเกณฑ์น้ำหนักน้อยหรือผอม
ค่า BMI 18.5 – 22.90	อยู่ในเกณฑ์ปกติ
ค่า BMI 23 – 24.90	น้ำหนักเกิน
ค่า BMI 25 – 29.90	โรคอ้วนระดับที่ 1
ค่า BMI 30 ขึ้นไป	โรคอ้วนระดับที่ 2

'sbp', = ค่าความดันตัวบน (Systolic blood pressure)

'dbp', = ค่าความดันตัวล่าง (Diastolic blood pressure)

'hb_val', = ค่าฮีโมโกบินส์ (Hemoglobin)

'fbs_val', = ระดับน้ำตาลในเลือด (Fasting Blood Sugar)+

'creatinine_val', = ระดับ creatinine ในเลือดเพื่อประเมินการทำงานของไต

'egfr_val', = ค่าคำนวนอัตราการคัดกรองของเสียของไต

'triglyceride_val', = ค่าไขมันไตรกลีเซอไรด์

't_cholesterol_val', = ค่าคลอเลสเตอรอล

'hdl_val', = ค่าไขมันดี

'ldl_val' = ค่าไขมันเลว

จากนั้นทำการลด outliers โดยใช้ bmi value เป็นตัวตัดค่า เพราะเนื่องจากภายในข้อมูลมีการลงค่าส่วนสูงกับน้ำหนักสลับกัน เราจึงจะไม่นำข้อมูลที่มี bmi value ผิดปกติมาใช้ในการให้แบบจำลองเรียนรู้และทำนายต่อไป

4.2 ผลการทดสอบประสิทธิภาพแบบจำลองพื้นฐาน

ในขั้นตอนของกระบวนการวิจัยนี้ คือการพัฒนาแบบจำลองสำหรับทำนายแนวโน้มการเปลี่ยนแปลงของราคาอ้างอิงในช่วงเวลาถัดไปโดยใช้ 3 เทคนิค ได้แก่ แบบจำลอง Naïve Bayes แบบจำลองเพื่อนบ้านใกล้เคียง k ตัว และแบบจำลองต้นไม้ตัดสินใน ซึ่งยังไม่ผ่านกระบวนการปรับปรุงประสิทธิภาพ และมีการเลือกจำนวน feature โดยการใช้กระบวนการ feature selection แบบ combining 3 model ตาม รูปที่ 3.5 ผล feature selection โดยการ combinig 3 model

4.2.1 model ที่ใช้ในการทำ features selection

4.2.1.1 Lasso Cross Validation

ในสถิติและการเรียนรู้ของเครื่อง lasso เป็นวิธีการวิเคราะห์การถดถอยที่ดำเนินการทั้งการเลือกตัวแปรและการทำให้เป็นมาตรฐานเพื่อเพิ่มความแม่นยำในการทำนายและการตีความของแบบจำลองทางสถิติที่เป็นผลลัพธ์

```
from sklearn.linear_model import LassoCV

# Create and fit the LassoCV model on the training set
lcv1 = LassoCV()
lcv1.fit(X1_train, y1_train)
print('Optimal alpha = {0:.3f}'.format(lcv1.alpha_))

# Calculate R squared on the test set
r_squared1 = lcv1.score(X1_test, y1_test)
print('The model explains {0:.1%} of the test set variance'.format(r_squared1))

# Create a mask for coefficients not equal to zero
lcv_mask1 = lcv1.coef_ != 0
print('{} features out of {} selected'.format(sum(lcv_mask1), len(lcv_mask1)))
print(X1_lcv_m1.columns)
```

```
Optimal alpha = 0.159
The model explains 17.7% of the test set variance
10 features out of 13 selected
Index(['bmi_val', 'sbp', 'dbp', 'hb_val', 'fbs_val', 'egfr_val',
       'triglyceride_val', 't_cholesterol_val', 'hdl_val', 'ldl_val'],
      dtype='object')
```

รูปที่ 4.3 model lasso feature selection แบบ cross validated

4.2.1.2 Recursive Feature Eliminator (RFE)

RFE เป็นวิธีการเลือกคุณลักษณะโดยอิงจาก Wrapper ซึ่งจะกำจัดคุณลักษณะแบบวนซ้ำโดยการฝึกแบบจำลองเกี่ยวกับชุดคุณลักษณะทั้งหมด และกำจัดคุณลักษณะที่จำเป็นน้อยที่สุดตามน้ำหนักหรือคะแนนนัยสำคัญ โดยจะค่อยๆ กำจัดคุณลักษณะต่างๆ ออกไปจนกว่าจะถึงจำนวนที่กำหนดไว้หรือเกณฑ์การหยุด โดยในการการทำ feature selection ที่ผู้วิจัยทำนี้จะใช้ RFE แบบ GradientBoostingRegressor กับ RFE แบบ RandomForestRegressor

```
from sklearn.feature_selection import RFE
from sklearn.ensemble import GradientBoostingRegressor

# Select 10 features with RFE on a GradientBoostingRegressor, drop 3 features on each step
rfe_gb1 = RFE(estimator=GradientBoostingRegressor(),
              n_features_to_select=10, step=3, verbose=1)
rfe_gb1.fit(X1_train, y1_train)

# Calculate the R squared on the test set
r_squared1 = rfe_gb1.score(X1_test, y1_test)
print('The model can explain {0:.1%} of the variance in the test set'.format(r_squared1))

# Assign the support array to gb_mask
gb_mask1 = rfe_gb1.support_

# Create a mask for coefficients not equal to zero
#lcv_mask1 = rfe_gb1.coef_ = ['True']
print('{} features out of {} selected'.format(sum(gb_mask1), len(gb_mask1)))
print(X1_gb_m1.columns)
```

```
Fitting estimator with 13 features.
The model can explain 90.0% of the variance in the test set
10 features out of 13 selected
Index(['bmi_val', 'sbp', 'hb_val', 'fbs_val', 'creatinine_val', 'egfr_val',
       'triglyceride_val', 't_cholesterol_val', 'hdl_val', 'ldl_val'],
      dtype='object')
```

รูปที่ 4.4 model Recursive Feature Eliminator แบบ GradientBoostingRegressor

```
from sklearn.feature_selection import RFE
from sklearn.ensemble import RandomForestRegressor

# Select 10 features with RFE on a RandomForestRegressor, drop 3 features on each step
rfe_rf1 = RFE(estimator=RandomForestRegressor(),
              n_features_to_select=10, step=3, verbose=1)
rfe_rf1.fit(X1_train, y1_train)

# Calculate the R squared on the test set
r_squared1 = rfe_rf1.score(X1_test, y1_test)
print('The model can explain {0:.1%} of the variance in the test set'.format(r_squared1))

# Assign the support array to rf_mask
rf_mask1 = rfe_rf1.support_

print('{} features out of {} selected'.format(sum(rf_mask1), len(rf_mask1)))
print(X1_rf_m1.columns)
```

```
Fitting estimator with 13 features.
The model can explain 92.6% of the variance in the test set
10 features out of 13 selected
Index(['bmi_val', 'sbp', 'dbp', 'hb_val', 'fbs_val', 'creatinine_val',
       'egfr_val', 't_cholesterol_val', 'hdl_val', 'ldl_val'],
      dtype='object')
```

รูปที่ 4.5 model Recursive Feature Eliminator แบบ RandomForestRegressor

ประสิทธิภาพด้านความแม่นยำในการทำนายกับแบบจำลองต่างๆ กับข้อมูลที่ยังไม่ผ่านกระบวนการใดๆ กับข้อมูลที่ผ่านการคัดเลือก feature แล้ว ซึ่งขั้นตอนการสร้างแบบจำลองพื้นฐานนี้ แสดงรายละเอียดดังถัดไป

4.2.2 แบบจำลองที่นำมาใช้ มีทั้งหมด 3 model

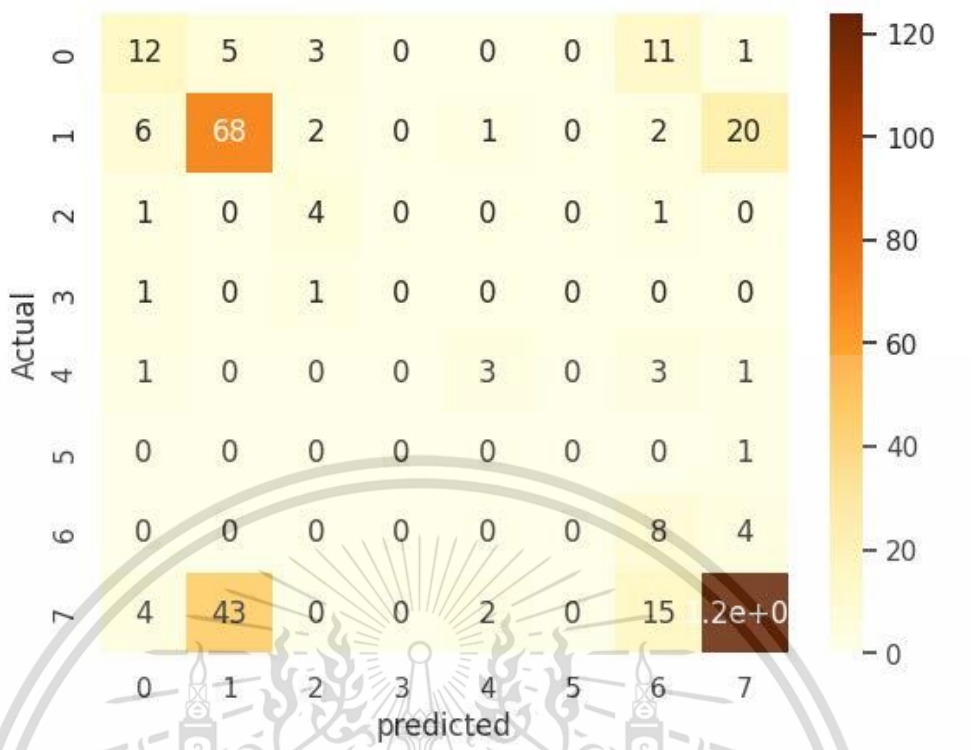
4.2.2.1 แบบจำลอง Naïve Bayes

แบบจำลอง Naive Bayes เป็นวิธีการเรียนรู้แบบมีผู้สอน supervised learning methods based โดยอาศัยการประยุกต์ใช้ทฤษฎีบทของเบย์ โดยหลักการของ NB จะใช้หลักการเรื่อง ความน่าจะเป็น (probability) ในการทำนายว่าเป็น กลุ่มไหน

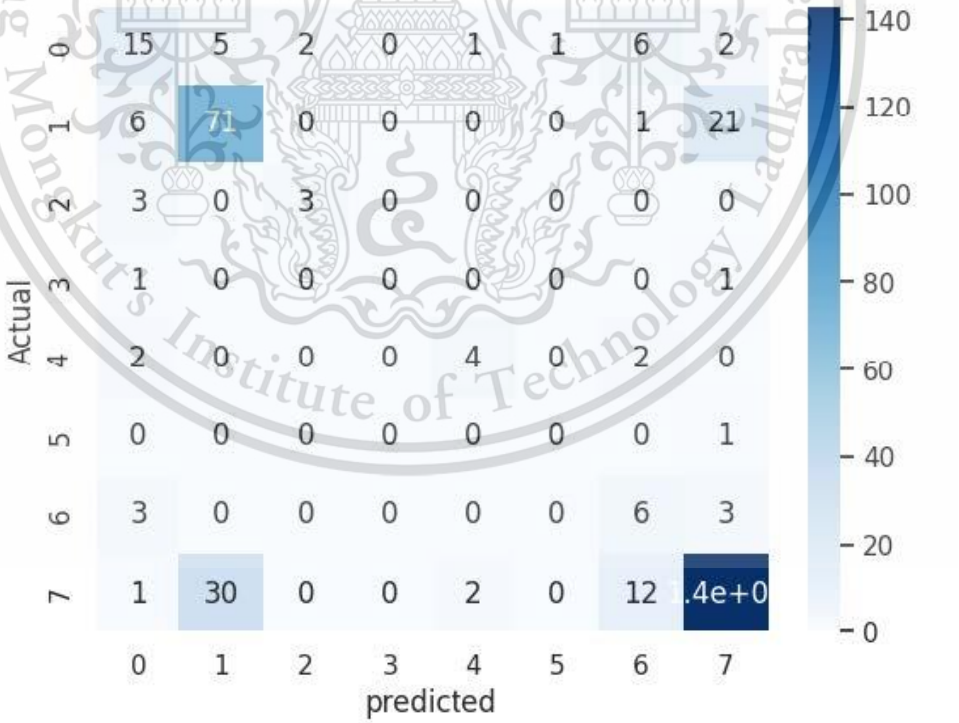
```
#from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()
nb_reduced = GaussianNB()
#y = volunteer['category_desc']

# Split the dataset according to the class distribution of category_desc,
# using the filtered_text vector
#X_train, X_test, y_train, y_test = train_test_split(filtered_text.toarray(), y, stratify=y)

# Fit the model to the training data
nb.fit(X1_train, y1_train)
nb_reduced.fit(X1_reduced_train, y1_reduced_train)

# Print out the model's accuracy
print('nb Accuracy : ',nb.score(X1_test, y1_test))
print('nb_reduced Accuracy : ',nb_reduced.score(X1_reduced_test, y1_reduced_test))
```

```
nb Accuracy :  0.6293103448275862
nb_reduced Accuracy :  0.6954022988505747
```

รูปที่ 4.6 ค่าความถูกต้องของ แบบจำลอง naïve bayes



รูปที่ 4.7 confusion metrix ของ แบบจำลอง naïve bayes ก่อนการทำ feature selection



รูปที่ 4.8 confusion metrix ของ แบบจำลอง naïve bayes หลังการทำ feature selection

```
1 from sklearn.metrics import classification_report
2 print('Naive Bayes before features selection calassification :=')
3 print(classification_report(y1_test, y1_pred_nb))
4 print('Naive Bayes after features selection calassification :=')
5 print(classification_report(y1_reduced_test, y1_pred_nb_reduced))
```

```
Naive Bayes before features selection calassification :=
              precision    recall  f1-score   support

           1       0.48      0.38      0.42        32
           2       0.59      0.69      0.63        99
           3       0.40      0.67      0.50         6
           4       0.00      0.00      0.00         2
           5       0.50      0.38      0.43         8
           6       0.00      0.00      0.00         1
           7       0.20      0.67      0.31        12
           8       0.82      0.66      0.73       188

    accuracy                           0.63       348
   macro avg       0.37      0.43      0.38       348
weighted avg       0.68      0.63      0.64       348

Naive Bayes after features selection calassification :=
              precision    recall  f1-score   support

           1       0.48      0.47      0.48        32
           2       0.67      0.72      0.69        99
           3       0.60      0.50      0.55         6
           4       0.00      0.00      0.00         2
           5       0.57      0.50      0.53         8
           6       0.00      0.00      0.00         1
           7       0.22      0.50      0.31        12
           8       0.84      0.76      0.80       188

    accuracy                           0.70       348
   macro avg       0.42      0.43      0.42       348
weighted avg       0.72      0.70      0.70       348
```

รูปที่ 4.9 classification report ของ แบบจำลอง naïve bayes ก่อนและหลังการทำ feature selection

4.2.2.2 แบบจำลอง K-Nearest Neighbors (KNN)

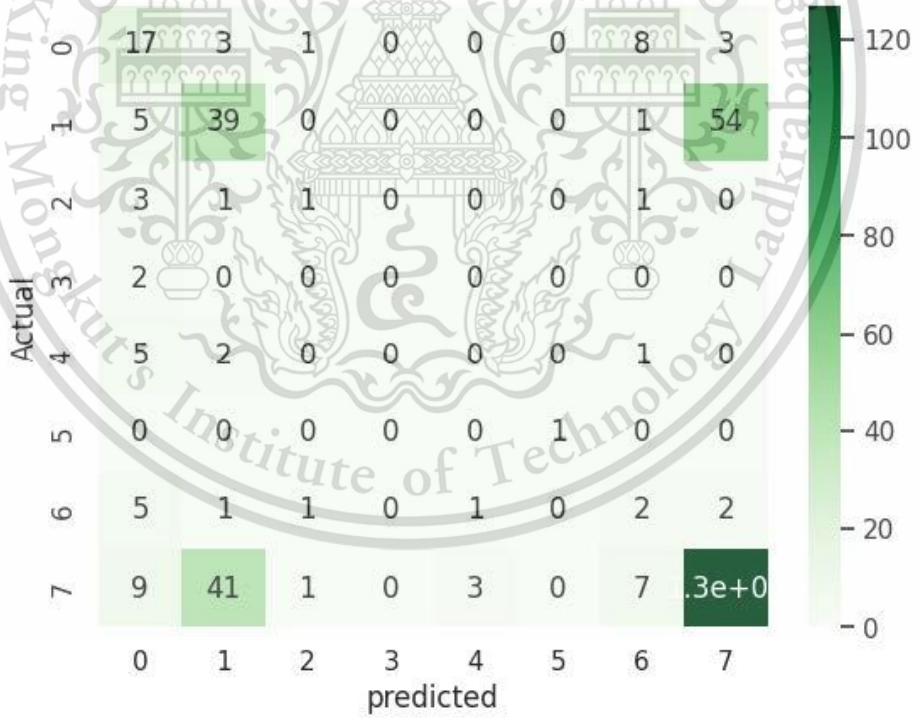
K-Nearest Neighbors หรือที่เรียกว่า KNN หรือ k-NN เป็นวิธีการแบ่งคลาสสำหรับใช้จัดหมวดหมู่ข้อมูล (Classification) โดยมีหลักการนำข้อมูลอื่นๆมาเปรียบเทียบกับตัวข้อมูลที่สนใจ ว่ามีความใกล้เคียงกันมากแค่ไหน หากข้อมูลที่สนใจอยู่ใกล้กับข้อมูลใดมากที่สุด ระบบจะให้คำตอบเป็นเหมือนคำตอบของข้อมูลที่อยู่ใกล้ที่สุด โดยส่วนใหญ่จะใช้สำหรับแก้ปัญหาที่เรารู้จำนวนกลุ่มที่แน่นอนอยู่แล้ว แต่มีบาง

ข้อมูลบางตัวที่เราไม่สามารถบอกได้ว่าข้อมูลนั้นอยู่กลุ่มไหน เราก็จะใช้ระบบนี้เข้ามาช่วยเลือก ซึ่งมันจะคล้ายๆกับการโหวตเสียงข้างมาก

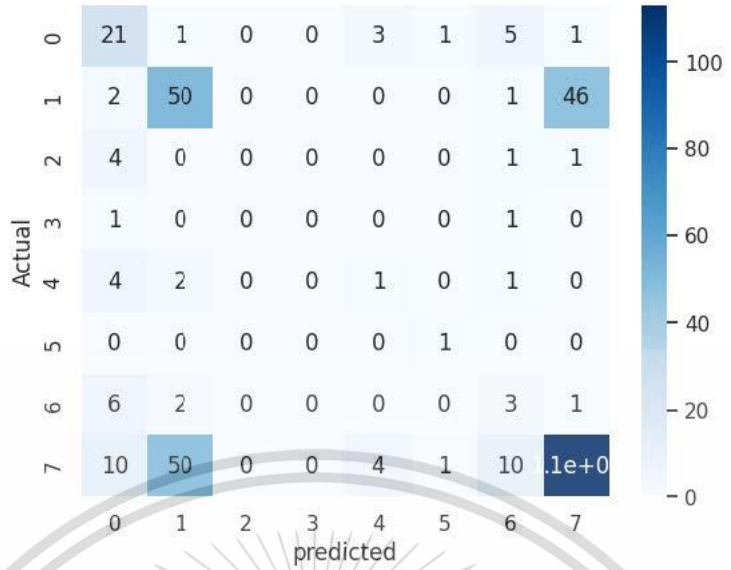
```
knn = KNeighborsClassifier(n_neighbors=5)
knn_rd = KNeighborsClassifier(n_neighbors=5)


# Fit knn to the training data
knn.fit(X1_train, y1_train)
knn_rd.fit(X1_reduced_train, y1_reduced_train)

# Score knn on the test data and print it out
print(knn.score(X1_test, y1_test))
print(knn_rd.score(X1_reduced_test, y1_reduced_test))
```

```
0.5373563218390804
0.5431034482758621
```

รูปที่ 4.10 ค่าความถูกต้องของ แบบจำลอง K-Nearest Neighbors (KNN)



รูปที่ 4.11 confusion metrix ของ แบบจำลอง K-Nearest Neighbors (KNN) ก่อนการทำ feature selection



รูปที่ 4.12 confusion metrix ของ แบบจำลอง K-Nearest Neighbors (KNN) หลังการทำ feature selection

```
1 #from sklearn.metrics import classification_report
2 print('knn before feature selection calassification :=')
3 print(classification_report(y1_test, y1_pred_knn))
4 print('knn after  feature selection calassification :=')
5 print(classification_report(y1_reduced_test, y1_reduced_pred_knn))
```

```
knn before feature selection calassification :=
              precision    recall  f1-score   support

           1       0.37      0.53      0.44        32
           2       0.45      0.39      0.42        99
           3       0.25      0.17      0.20         6
           4       0.00      0.00      0.00         2
           5       0.00      0.00      0.00         8
           6       1.00      1.00      1.00         1
           7       0.10      0.17      0.12        12
           8       0.68      0.68      0.68       188

    accuracy                           0.54       348
   macro avg       0.36      0.37      0.36       348
weighted avg       0.54      0.54      0.54       348

knn after  feature selection calassification :=
              precision    recall  f1-score   support

           1       0.44      0.66      0.53        32
           2       0.48      0.51      0.49        99
           3       0.00      0.00      0.00         6
           4       0.00      0.00      0.00         2
           5       0.12      0.12      0.12         8
           6       0.33      1.00      0.50         1
           7       0.14      0.25      0.18        12
           8       0.70      0.60      0.65       188

    accuracy                           0.54       348
   macro avg       0.28      0.39      0.31       348
weighted avg       0.56      0.54      0.55       348
```

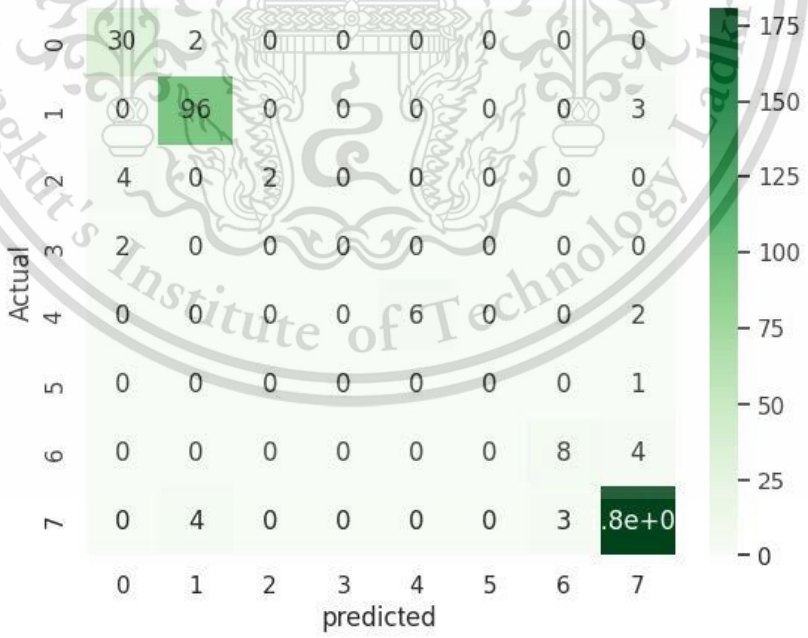
รูปที่ 4.13 classification report ของ แบบจำลอง k-nearest neighbors ก่อนและหลังการทำ feature selection

4.2.2.3 แบบจำลอง Decision Tree

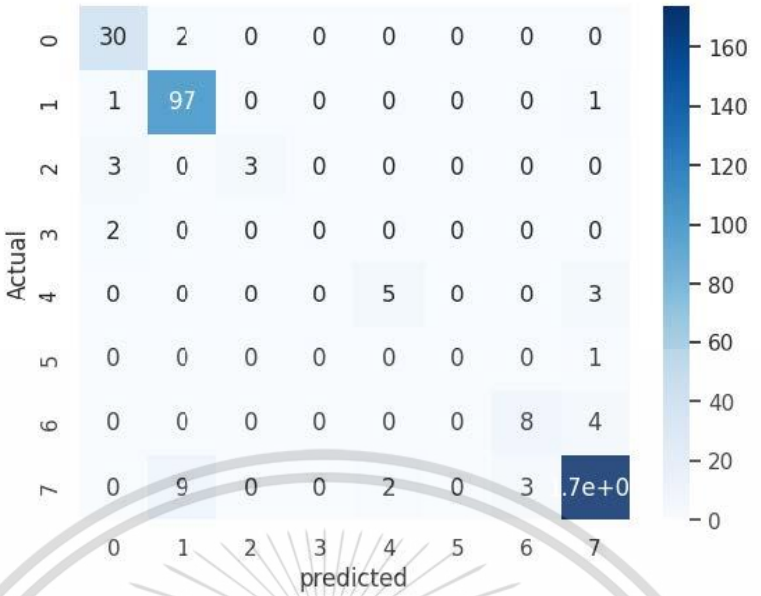
Decision Tree หรือ ต้นไม้ตัดสินใจ คือ การจำลองวิธีการตัดสินใจของมนุษย์ ซึ่งการตัดสินใจแต่ละครั้งของมนุษย์ มนุษย์จะแตกโจทย์หลักออกเป็นโจทย์ย่อยหลาย ๆ โจทย์ก่อนเพื่อที่จะได้ง่ายต่อการตัดสินใจ หรือจะนำเอาปัจจัยต่าง ๆ ที่เกี่ยวข้องกับการตัดสินใจหรือเกี่ยวข้องกับโจทย์หลักมาตั้งเป็นคำถามใหม่หรือแตกเป็นโจทย์ย่อย และถามตัวเองใหม่อีกครั้ง มีโครงสร้างแบบต้นไม้เป็นลำดับชั้น ซึ่งประกอบด้วยโหนดรูท กิ่งก้าน โหนดภายใน และโหนดใบ

```
1 clf_dt1 = RandomForestClassifier(random_state=421)
2 clf_dt11 = clf_dt1.fit(X1_train, y1_train)
3 print('dt Accuracy := ',clf_dt11.score(X1_test,y1_test))
4
5 clf_dt2 = RandomForestClassifier(random_state=421)
6 clf_dt21 = clf_dt2.fit(X1_reduced_train, y1_reduced_train)
7 print('dt reduced Accuracy := ',clf_dt21.score(X1_reduced_test,y1_reduced_test))
8 #fig = plt.figure(figsize=(80,45))
9 #visualize_classifier(clf_dt1, X1_test, y1_test);

dt Accuracy := 0.9281609195402298
dt reduced Accuracy := 0.9109195402298851
```

รูปที่ 4.14 ค่าความถูกต้องของ แบบจำลอง Decision Tree



รูปที่ 4.15 confusion metrix ของ แบบจำลอง Decision Tree ก่อนการทำ feature selection



รูปที่ 4.16 confusion metrix ของ แบบจำลอง Decision Tree หลังการทำ feature selection

decision tree before features selection calassification :=

	precision	recall	f1-score	support
1	0.83	0.94	0.88	32
2	0.94	0.97	0.96	99
3	1.00	0.33	0.50	6
4	0.00	0.00	0.00	2
5	1.00	0.75	0.86	8
6	0.00	0.00	0.00	1
7	0.73	0.67	0.70	12
8	0.95	0.96	0.96	188
accuracy			0.93	348
macro avg	0.68	0.58	0.61	348
weighted avg	0.92	0.93	0.92	348

decision tree after features selection calassification :=

	precision	recall	f1-score	support
1	0.83	0.94	0.88	32
2	0.90	0.98	0.94	99
3	1.00	0.50	0.67	6
4	0.00	0.00	0.00	2
5	0.71	0.62	0.67	8
6	0.00	0.00	0.00	1
7	0.73	0.67	0.70	12
8	0.95	0.93	0.94	188
accuracy			0.91	348
macro avg	0.64	0.58	0.60	348
weighted avg	0.90	0.91	0.91	348

รูปที่ 4.17 classification report ของ แบบจำลอง k-nearest neighbors ก่อนและหลังการทำ feature selection

4.3 อภิปรายผลการวิจัย

วัตถุประสงค์หลักของงานวิจัยนี้คือการพัฒนาเครื่องมือเพื่อช่วยให้เป็นแนวทางการวินิจฉัยโรคที่ได้จากค่า lab ในการตรวจสุขภาพ ซึ่งจากการทำการทดลอง ได้นำเสนอขั้นตอนการเตรียมข้อมูลและกระบวนการที่ใช้ในการพัฒนา โดยเลือกใช้ ทฤษฎีการเรียนรู้ของเครื่อง ได้แก่ แบบจำลอง Naïve Bayes แบบจำลองเพื่อนบ้านใกล้เคียง k ตัว และแบบจำลองต้นไม้ตัดสินใจ จากผลลัพท์ที่ได้จากการทดสอบทำให้สรุปเป็นประเด็นต่างๆ ได้ดังนี้

1) ผลสรุปที่ได้จากการทดสอบประสิทธิภาพทั้ง 3 รูปแบบ โดยกระบวนการพื้นฐาน และยังไม่ผ่าการทำ features selection พบว่าแบบจำลองมีความแม่นยำในการทำนาย โดยแบบจำลอง Naïve Bayes มีค่าเท่ากับ 0.6293 แบบจำลองเพื่อนบ้านใกล้เคียง k ตัวมีค่าเท่ากับ 0.5373 และแบบจำลองต้นไม้ตัดสินใจมีค่าเท่ากับ 0.9321
2) ผลสรุปที่ได้จากการทดสอบประสิทธิภาพทั้ง 3 รูปแบบ โดยกระบวนการพื้นฐาน และผ่านการทำ features selection พบว่าแบบจำลองมีความแม่นยำในการทำนาย โดยแบบจำลอง Naïve Bayes มีค่าเท่ากับ 0.6954 แบบจำลองเพื่อนบ้านใกล้เคียง k ตัวมีค่าเท่ากับ 0.5431 และแบบจำลองต้นไม้ตัดสินใจมีค่าเท่ากับ 0.8938
3) จากผลการทดสอบโดยวิธีพื้นฐานโดยที่ไม่ผ่าน และผ่านกระบวนการ feature selection แบบจำลองต้นไม้ตัดสินใจมีค่าความแม่นยำมากที่สุดคือ 0.9321 และ 0.8938 ตามลำดับ ทำให้การจัดกลุ่มได้ถูกต้องมากที่สุดใน 3 รูปแบบจำลอง

บทที่ 5
สรุปผลงานวิจัยและข้อเสนอแนะ

งานวิจัยนี้ได้นำความรู้ทางด้านปัญญาประดิษฐ์ โดยเฉพาะเทคนิคการเรียนรู้ของเครื่อง ได้แก่ แบบจำลอง Naïve Bayes, แบบจำลองเพื่อนบ้านใกล้เคียง k ตัว และแบบจำลองต้นไม้ตัดสินใจ มาประยุกต์เพื่อพัฒนาเป็นแบบจำลองสำหรับการทำนายโรคจากผลการตรวจสุขภาพ ซึ่งแบบจำลองนี้จะช่วยให้แพทย์ หรือผู้ป่วย กำหนดทิศทางในการดำรงชีวิต เพื่อรักษาสุขภาพของผู้ที่มารับการตรวจสุขภาพ นอกจากนี้ยังได้นำเสนอวิธีการต่างๆ เช่นการรับประทานอาหาร ในด้านของการสร้างเสริมสุขภาพของผู้ที่ได้รับการตรวจสุขภาพ จากผลการศึกษาทั้งหมด สามารถสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลงานวิจัย

ผลสรุปจากการทำการทดลอง สรุปว่า แบบจำลอง Naïve Bayes และ K-NN นั้น ให้ค่าความแม่นยำดีขึ้นเล็กน้อยหลังจากได้ผ่านการทำ feature selection แต่แบบจำลองต้นไม้ตัดสินใจกลับได้รับผลแย่ลงเล็กน้อย เมื่อเทียบระหว่างก่อน และหลังการทำ feature selection โดยค่าความถูกต้องที่ดีที่สุดของก่อนและหลัง การทำ features selection ในแบบจำลอง Naïve Bayes มีค่าเท่ากับ 0.6954, แบบจำลองเพื่อนบ้านใกล้กัน k ตัว มีค่าเท่ากับ 0.5431 และแบบจำลองต้นไม้ตัดสินใจมีค่าเท่ากับ 0.9321

หลังจากทำการฝึกฝนแบบจำลองแล้ว นำมาทำนายว่าความถูกต้องของแบบจำลอง ปรากฏว่าค่า ไขมันผิดปกติมีค่ามากเป็นที่ 2 โดยค่ามากที่ 1 ได้แก่ มีค่าผิดปกติมากกว่า 2 ค่า ซึ่งค่าความถูกต้องตามแบบ ที่ได้มีการอ้างอิงไว้ใน 4.3 การอภิปรายผลวิจัย ข้อ 1 และ 2 เพราะฉะนั้นจากผลกาทดลองควรใช้ แบบจำลองต้นไม้ตัดสินใจในการนำมาทำนายโดยไม่จำเป็นต้องผ่านกระบวนการทำ features selection มาก่อนซึ่งมีค่าความถูกต้อง 93.21%

ตารางที่ 5.1 จำนวนข้อมูลของ class ใน dataset

class	class_label	จำนวน
8	มีความผิดปกติมากกว่า 1 ค่า	758
2	ไขมันผิดปกติ	295
1	ปกติ	96
7	BMI ผิดปกติ	85
5	ระดับฮีโมโกลบินผิดปกติ	28
3	ไตผิดปกติ	17
6	ความดันผิดปกติ	4
4	น้ำตาลผิดปกติ	2

5.2 ข้อเสนอแนะ

จากผลการศึกษาที่ได้ในงานวิจัยนี้ สามารถสรุปเป็นประเด็นต่างๆ ที่เป็นการเสนอแนะเพื่อการต่อยอดงานวิจัยได้ดังต่อไปนี้

1) ในงานวิจัยนี้ผู้วิจัยใช้ข้อมูลการตรวจสุขภาพประจำปีของพนักงานสถาบันการแพทย์จักรีนฤบดินทร์ และโรงพยาบาลรามาธิบดีจักรีนฤบดินทร์ ซึ่งเฉลี่ยอายุของพนักงานจะไม่เกิน 35 ปี โดยมีสัดส่วนถึง 80% ของข้อมูลที่นำมาใช้งาน ทำให้การแปลผลนี้สามารถวัดผลได้เฉพาะพนักงานที่ทำงานอยู่ องค์กรณ์นี้เท่านั้น
2) ตัวแบบจำลองที่นำมาใช้ในการทำวิจัยนี้จะสังเกตว่า ให้ค่าความถูกต้อง (Accuracy) อยู่ในค่า 0.5 – 0.6 ในแบบจำลองต้นไม้ตัดสินใจ และแบบทฤษฎี Naïve Bayes แต่ถ้านำตัวแบบจำลองนี้ ไปใช้ใน dataset อื่นผลที่ได้อาจจะดีกว่านี้ เพราะข้อมูลที่ใช้ในการวิจัยนี้มีค่าเฉลี่ยอายุอยู่ที่ไม่เกิน 35 ปี
3) ในการทำวิจัยนี้เราอาจจะเพิ่มประสิทธิภาพของ model ให้มากขึ้นโดยการนำข้อมูลมาทำการ imbalance data โดการ Upsampling หรือ Downsampling ข้อมูลของเรา แต่วิธี 2 วิธีจะเป็นการเพิ่มข้อมูลของเราขึ้นมา (อย่างมีหลักการ) เนื่องจากข้อมูล ของ raw data ที่นำมาทำนายใน class ต่างๆ มีไม่เท่ากันก็จะทำให้ model มีประสิทธิภาพมากขึ้น
4) ในขั้นตอนการกำหนดคลาสเป้าหมายให้แก่ข้อมูล ผู้วิจัยได้กำหนดเพียง 8 class โดยกำหนดแค่ว่า ปกติ ไขมันผิดปกติ น้ำตาลในเลือดผิดปกติ ค่าฮีโมโกลบิน ค่าไตผิดปกติ ความดันผิดปกติ ค่า BMI ผิดปกติ และมีความผิดปกติมากกว่า 1 ค่า ถ้าแบ่ง class ที่มีความน่าจะเป็น ไปได้ จะได้เท่ากับ 2^6 เท่ากับ 64 +1 class ถ้าต้องการให้ผลทำนายมีความถูกต้องตามโรค ที่น่าจะเกิดขึ้นจริงได้
5) แม้นว่าการวิจัยนี้มีเป้าหมายเพื่อพัฒนาเครื่องมือ ในการทำนายการเกิดโรคที่มาจากการตรวจสุขภาพประจำปี แต่ก็สามารถนำมาประยุกต์ใช้กับการทำนายแบบต่างๆ ที่ใช้การจัดกลุ่มได้ โดยอาจเปลี่ยนตัวแปรเพื่อนำมาทำนายได้
6) ถึงแม้นว่าการวิจัยนี้จัดทำเพื่อการทำนายโรค แต่อาจจะเป็นได้แค่ guild line เพื่อช่วยให้แพทย์นำมาประกอบการวินิจฉัยโรค ในแต่ละบุคคล เพราะฉนั้นท้ายที่สุดควรให้แพทย์เป็นผู้วินิจฉัยสุดท้าย เพื่อความถูกต้องตามหลักการรักษา
7) จากการทดลอง จะเห็นได้ว่าค่าไขมันในจำนวนตัวอย่างการทดลอง มีความผิดปกติมาก จึงทำให้ผู้บริหารองค์กรณ์ สามารถจัดกิจกรรมให้พนักงานในการลดไขมัน เพื่อสุขภาพของพนักงาน เป็นต้น

เอกสารอ้างอิง (References)

[1] RolandDominic G, Jamora, MD, PhD Mario B. Prado Jr., MD Veeda Michelle M. Anlacan, MD Marie Charmaine C. Sy, MD, MBA and Adrian I. Espiritu, MD (2022). Incidence and riskfactors for stroke in patients with COVID-19 in the Philippines: Ananalysis of 10,881 cases. Journal of Stroke and Cerebrovascular Diseases, Vol.31, No.11 (November), 2022:106776

[2] Duen-Yian Yeh, Ching-Hsue Cheng, Yen-Wen Chen (2011). A predictive model for cerebrovascular disease using data mining. Expert Systems with Applications 38 (2011) 8970–8977

[3] José Alberto Tavares Rodríguez (2021). Stroke prediction through Data Science and Machine Learning Algorithms. Preprint · June 2021 DOI:0.13140/RG.2.2.33027.43040

[4] Naruemol Kingkaew and Tidarat Antadech (2019). Cardiovascular risk factors and 10-year CV risk scores in adults aged 30-70 years old in Amnat Charoen Province, Thailand. Asia-Pacific Journal of Science and Technology: Volume: 24. Issue: 04. Article ID.: APST-24-04-04.

[5] คณะแพทย์ศาสตร์ศิริราชพยาบาล (2566). ดัชนีมวลกาย สำคัญอย่างไร. สืบค้นเมื่อ 10 พฤศจิกายน 2566, จาก https://www.si.mahidol.ac.th/th/healthdetail.asp?aid=1361

[6] โรงพยาบาลเมดพาร์ด (2566). ไขมันในเลือดสูง ไขมันในเลือดแต่ละชนิดต่างกันอย่างไร. สืบค้นเมื่อ 10 พฤศจิกายน 2566, จาก http://www.medparkhospital.com/disease-and-treatment/how-blood-lipids-different-from-each-other

[7] โรงพยาบาลศิริราช ปิยมหาราชการุณย์ (2566). ไขมันในเลือดสูงเสี่ยงโรคหัวใจและหลอดเลือด. สืบค้นเมื่อ 10 พฤศจิกายน 2566, จาก https://www.siphhospital.com/th/news/article/share/dyslipidemia

[8] โรงพยาบาลวิมุต (2566). ภาวะน้ำตาลในเลือดสูงเท่าไหร่เสี่ยงเป็นเบาหวาน. สืบค้นเมื่อ 10 พฤศจิกายน 2566, จาก https://www.vimut.com/article/hyperglycemia-symptoms

[9] โรงพยาบาลศิริราช ปิยมหาราชการุณย์ (2566). โรคไตเรื้อรังไม่ยากเข้าใจ. สืบค้นเมื่อ 10 พฤศจิกายน 2566, จาก https://www.siphhospital.com/th/news/article/share/461

[10] กรมอนามัย (2566). กรมอนามัย เผยไทยติด 1 ใน 5 ประเทศที่มีอัตราการเกิดโรคไตสูงสุด แนะเลี่ยง 8 ประเภทอาหาร. สืบค้นเมื่อ 10 พฤศจิกายน 2566, จาก https://multimedia.anamai.moph.go.th/news/090366/

เอกสารอ้างอิง (ต่อ)

[11] กรมควบคุมโรค (2565). สถิติโรคความดันโลหิตสูง, ก - หน้าแรก | กรมควบคุมโรค – กระทรวงสาธารณสุข. เบาหวาน. สืบค้นเมื่อ 10 พฤศจิกายน 2566, จาก https://ddc.moph.go.th/brc/news.php?news=25290&deptcode=brc&news_views=388

[12] โรงพยาบาลศิริราช ปิยมหาราชการุณย์ (2566). คุณเป็นความดันโลหิตสูงรึเปล่า?. สืบค้นเมื่อ 10 พฤศจิกายน 2566, จาก https://www.siphhospital.com/th/news/article/share/hypertension

[13] Peachapong Poolpol (2566). การทำนายค่าระดับน้ำตาลในเลือด จากผลข้อมูลการตรวจสุขภาพ ในเจ้าหน้าที่โรงพยาบาลแห่งหนึ่ง โดยใช้โมเดล Linear Regression. สืบค้นเมื่อวันที่ 10 พฤศจิกายน 2566, จาก https://peachapong-poolpol.medium.com/การทำนายค่าระดับน้ำตาลในเลือด-จากผลข้อมูลการตรวจสุขภาพ-ในเจ้าหน้าที่โรงพยาบาลแห่งหนึ่ง-โดยใช้โมเดล-29afba9a678e

[14] google colab (2566). Feature selection II - selecting for model accuracy. สืบค้นเมื่อวันที่ 10 พฤศจิกายน 2566, จาก https://colab.research.google.com/github/goodboychan/chans_jupyter/blob/main/_notebooks/2020-07-08-03-Feature-selection-II-selecting-for-model-accuracy.ipynb

ประวัติผู้เขียน

ชื่อ	นาย มงคล ตปนียปทุม
วัน เดือน ปี เกิด	19 มกราคม 2519
ที่อยู่ปัจจุบัน	212/58 หมู่บ้านพฤกษาวิลล์ 34 ซอยเพชรเกษม 110 แขวงหนองค้างพลู เขตหนองแขม กรุงเทพฯ 10160
ประวัติการศึกษา	(2542) วิทยาศาสตร์บัณฑิต วิทยาการคอมพิวเตอร์ เกรดเฉลี่ย 2.57 (ราชาภัฏธนบุรี)
ทุนการศึกษาที่ได้รับ	ไม่มี
อาชีพปัจจุบัน	นักวิชาการคอมพิวเตอร์ สถาบันการแพทย์จักรีนฤบดินทร์