ระบบแนะนำวัคซีนสำหรับคลินิกเด็กสุขภาพดี

สิริวรรณ พงษศิริ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการสารสนเทศ
คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
2564

ระบบแนะนำวัคซีนสำหรับคลินิกเด็กสุขภาพดี

สิริวรรณ พงษศิริ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการสารสนเทศ
คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
2564

Vaccine Recommendation System for Well Baby Clinic

SIRIWAN PHONGSASIRI

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE MASTER DEGREE OF SCIENCE

IN INFORMATICS

FACULTY OF INFORMATICS

BURAPHA UNIVERSITY

2021

คณะกรรมการควบคุมวิทยานิพนธ์และคณะกรรมการสอบวิทยานิพนธ์ได้พิจารณาวิทยานิพนธ์ของ สิริวรรณ พงษ์ศิริ ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการสารสนเทศ ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมวิทยานิพนธ์

อาจารย์ที่ปรึกษาหลัก

.....

(ผู้ช่วยศาสตราจารย์ ดร.สุวรรณา รัศมีขวัญ)

คณะกรรมการสอบวิทยานิพนธ์

_____ ประธาน

(รองศาสตราจารย์ ดร.อัณณ์นุพันธ์ รอดทุกข์)

_____ กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร)

_____ กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.สุวรรณา รัศมีขวัญ)

_____ คณบดีคณะวิทยาการสารสนเทศ

(ผู้ช่วยศาสตราจารย์ ดร. กฤษณะ ชินสาร)

วันที่_____เดือน_____พ.ศ._____

บัณฑิตวิทยาลัย มหาวิทยาลยบูรพา อนุมัติให้รับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการสารสนเทศ ของมหาวิทยาลัยบูรพา

_____ คณบดีบัณฑิตวิทยาลัย

(รองศาสตราจารย์ ดร.นุจรี ไชยมงคล)

วันที่ 8 เดือน กุมภาพันธ์ พ.ศ. 2564

61910138: สาขาวิชา: วิทยาการสารสนเทศ; วท.ม. (วิทยาการสารสนเทศ)
คำสำคัญ: ระบบแนะนำ, การเรียนรู้ของเครื่องจักร, Ensemble Learning

สิริวรรณ พงษศิริ : ระบบแนะนำวัคซีนสำหรับคลินิกเด็กสุขภาพดี. (Vaccine Recommendation System for Well Baby Clinic) คณะกรรมการควบคุมวิทยานิพนธ์: สุวรรณา รัศมีขวัญ ปี พ.ศ. 2564.

ในงานวิจัยนี้ ได้นำเสนอระบบแนะนำวัคซีนสำหรับคลินิกเด็กสุขภาพดี ซึ่งกรอบการทำงานของ แบ่งเป็น 2 เฟส คือ เฟสที่ 1 จะเป็นการจัดการข้อมูล โดยมีการทำความสะอาดข้อมูลและเติมข้อมูลด้วยค่าเฉลี่ย มีการกำจัดข้อมูลที่ผิด (outliers) ด้วยขั้นตอนวิธี Probabilistic Mapped Mean-Shift (PMMS) ซึ่งมีค่าความถูกเท่ากับ 93%, 94%, 80%, 75%, และ 72% เมื่อนำไปทดลองกับข้อมูล CWC, Stamps, Arrh, Pima และ Pakinson ตามลำดับ โดยค่าความถูกต้องที่ได้ดังกล่าวนี้เป็นค่าความถูกต้องสูงสุด เมื่อเทียบกับขั้นตอนวิธีอื่น ที่ใช้ในการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอ เมื่อจัดการกับข้อมูลในเฟสที่ 1 เรียบร้อยแล้ว ทำให้ได้ชุดข้อมูล CWC ที่ผ่านการทำความสะอาดและมีการเติมข้อมูล รวมถึงได้มีการกำจัดข้อมูลผิดปกติออกไป (cleaned and completed dataset without outliers) ก็จะนำข้อมูล CWC ไปทดลองเพื่อหาขั้นตอนวิธีที่เหมาะสมในการแนะนำวัคซีนรายคนสำหรับเด็ก ในเฟสที่ 2 ซึ่งพบว่า ขั้นตอนวิธี Gradient Boosting Classifier ให้ค่าความถูกต้องสูงสุดอยู่ที่ 53% ซึ่งเป็นค่าที่สูงสุดจากขั้นตอนวิธีทั้งหมด 11 วิธี

61910138: MAJOR: INFORMATICS; M.Sc. (INFORMATICS)
KEYWORDS: Recommendation System, Machine Learning, Ensemble Learning

SIRIWAN PHONGSASIRI : VACCINE RECOMMENDATION SYSTEM FOR WELL BABY CLINIC. ADVISORY COMMITTEE: SUWANNA RASMEQUAN, Ph.D. 2021.

In this research has introduced a Vaccine Recommendation system for Well Baby Clinic. The framework is divided into 2 phases: Phase 1 will be data management. The data was cleaned and filled with averages. Outliers are eliminated through algorithms. Probabilistic Mapped Mean-Shift (PMMS) with 93%, 94%, 80%, 75%, and 72% accuracy was tested with CWC, Stamps, Arrh, Pima and Pakinson data, respectively. This is the highest accuracy compared to other algorithms used to compare the performance of the proposed algorithm. Once the data in Phase 1 has been dealt with, This results in a clean and populated CWC dataset. Including cleaned and completed dataset without outliers, the CWC data will be tested to determine an appropriate algorithm for recommending individual vaccines for children. The Gradient Boosting Classifier method yields a maximum accuracy of 53%, which is the highest of 11 algorithms.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้คงไม่อาจสำเร็จสมบูรณ์ขึ้นมาได้หากปราศจากความเมตตา กรุณาจากท่านผู้ช่วยศาสตราจารย์ดร. สุวรรณา รัศมีขวัญ ที่ทำให้ผู้เขียนได้หัวข้อในการทำวิทยานิพนธ์และกรุณารับเป็นอาจารย์ที่ปรึกษาของผู้เขียน ได้ให้ข้อมูลและคำแนะนำต่าง ๆ ซึ่งเป็นประโยชน์ต่อผู้เขียน โดยเฉพาะการวางเค้าโครง แนวทางการทำวิจัยตลอดจนการกำหนดกรอบเวลาในการเสนอความคืบหน้าของงาน ซึ่งเป็นแรงกระตุ้นให้แก่ผู้เขียนได้อย่างดี ทั้งท่านอาจารย์ได้สละเวลาอันมีค่าตรวจสอบความถูกต้องของงานผู้เขียนอีกด้วย ผู้เขียนรู้สึกซาบซึ้งใจและสำนึกในพระคุณของท่านอาจารย์เป็นอย่างยิ่ง จึงขอกราบขอบพระคุณท่านอาจารย์ไว้ณ ที่นี้

ผู้เขียนกราบขอบคุณอาจารย์ รองศาสตราจารย์ ดร. อัณณ์นุพันธ์ รอดทุกข์ ประธานกรรมการวิทยานิพนธ์ อาจารย์ผู้ช่วยศาสตราจารย์ดร. กฤษณะ ชินสาร และ ผู้ช่วยศาสตราจารย์ดร.สุวรรณา รัศมีขวัญ กรรมการวิทยานิพนธ์ที่ท่านได้กรุณาชี้แนะแนวทางและคำแนะนำ จนทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลงได้

ผู้เขียนขอขอบคุณกัลยาณมิตรของผู้เขียนทุกท่านที่ให้ความช่วยเหลือและเป็นกำลังใจให้เสมอมา ขอบคุณพี่ๆ เพื่อนๆ และน้อง ทุกท่านที่ให้กำลังใจและคอยช่วยเหลือตลอดระยะเวลาที่ศึกษาและทำวิทยานิพนธ์เสมอมา

สุดท้ายผู้เขียนขอกราบขอบคุณ คุณพ่อ คุณแม่ คุณอา และพี่ชาย ที่คอยสนับสนุนในด้านการศึกษาตลอดมา เป็นกำลังใจที่สำคัญ ทำให้วิทยานิพนธ์สำเร็จลุล่วงได้ หากวิทยานิพนธ์ฉบับนี้มีความผิดพลาดประการใด ผู้เขียนขอน้อมรับผิดเพียงผู้เดียว

สิริวรรณ พงษศิริ

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญ

ในปัจจุบันการจัดการข้อมูลในโรงพยาบาล ได้อาศัยเทคโนโลยีคอมพิวเตอร์มาช่วยในการดำเนินการตั้งแต่การบันทึกข้อมูลประวัติผู้ป่วย ประวัติการรักษา และประวัติการรับยาเป็นต้น ซึ่งข้อมูลเหล่านี้เมื่อมีปริมาณมากขึ้น จะทำให้โรงพยาบาลมีข้อมูลมากเพียงพอที่จะสามารถนำมาใช้ในการวิเคราะห์ แต่การที่จะทำให้ผลการวิเคราะห์เกิดประโยชน์ต่อโรงพยาบาล ข้อมูลที่ถูกบันทึกต้องมีความครบถ้วนและถูกต้อง ก่อนที่จะสามารถนำไปใช้ประโยชน์ได้อย่างเหมาะสม อาทิ อาจนำไปใช้ในการแนะนำวัคซีนแต่ละชนิด ที่จำเป็นต้องใช้ในแต่ละช่วงเวลา โดยการคาดการณ์จาก แผนการมารับวัคซีนของเด็กที่เป็นไปตามกำหนดเวลา ซึ่งหากทำได้ ก็จะสามารถช่วยในการวางแผนและเตรียมการที่เกี่ยวข้องในการบริหารการจัดการคลังยา เพื่อเตรียมสั่งยาที่คาดว่าจะต้องใช้ได้ล่วงหน้าในปริมาณที่เหมาะสม โดยอาจมีการพยากรณ์ปริมาณยาที่จำเป็นต้องใช้ในแต่ละรายการ ซึ่งเป็นผลจากการวิเคราะห์ข้อมูลการเข้ารับวัคซีน และปริมาณยาที่จำเป็นต้องได้รับ เพื่อช่วยให้โรงพยาบาลสามารถบริหารจัดการคลังยาได้อย่างมีประสิทธิภาพ

สำหรับคลินิกเด็กสุขภาพดี ซึ่งให้บริการฉีดวัคซีนสำหรับเด็กอายุระหว่าง แรกเกิดจนถึง 6 ปี ในการให้บริการฉีดวัคซีนสำหรับเด็ก ณ คลินิกเด็กสุขภาพดีนี้ ทีมบุคลากรทางการแพทย์ที่ให้บริการ จะเริ่มต้นจากการตรวจสอบประวัติคนไข้ เช่น อายุ เพศ ประวัติการเคยได้รับวัคซีน เป็นต้น แล้วจึงฉีดวัคซีนให้กับเด็กตามช่วงอายุ และพัฒนาการของเด็ก และทำการบันทึกประวัติการให้วัคซีน เมื่อให้บริการแล้วเสร็จ ซึ่งแพทย์ผู้ดูแลคลินิก ได้ให้ข้อมูลว่า การให้วัคซีนกับเด็กนั้น ในกรณีที่เด็กมีประวัติการรักษาที่คลินิก และมารับวัคซีนตามระยะเวลา การให้วัคซีนตามช่วงอายุและพัฒนาการของเด็ก สามารถทำได้โดยสะดวก แต่ในกรณีที่เด็กไม่มีประวัติที่คลินิก หรือมีประวัติแต่ไม่มาเข้ารับการรักษาตามระยะเวลา จะทำให้แพทย์ต้องมีขั้นตอนและใช้เวลาในการพิจารณาวัคซีนสำหรับเด็กรายนั้น ๆ มากขึ้น เพื่อจะหาวัคซีนตัวทดแทนที่เหมาะสมกับช่วงอายุและพัฒนาการทางร่างการของเด็ก นอกจากประเด็นของกระบวนการให้บริการ ในกรณีที่เด็กมารับวัคซีนแบบไม่ต่อเนื่อง ที่สร้างความยุ่งยากในการให้บริการ ยังมีประเด็นการบันทึกข้อมูลเข้าสู่ระบบฐานข้อมูล ซึ่งผู้ช่วยแพทย์ไม่

สามารถบันทึกข้อมูลเข้าสู่ระบบได้ทันที เนื่องจากมีผู้มารับบริการจำนวนมาก จึงต้องจดบันทึกไว้ก่อนแล้วมาบันทึกข้อมูลภายหลัง ซึ่งในหลาย ๆ ครั้ง ก็พบว่า การบันทึกข้อมูลอาจไม่สมบูรณ์ หรืออาจเกิดข้อผิดพลาดในการบันทึกข้อมูล

จากความสำคัญของปัญหาข้างต้น ในงานวิจัยนี้จึงนำเสนอ ระบบแนะนำวัคซีนสำหรับคลินิกเด็กสุขภาพดี โดยเป็นการนำประวัติของเด็ก เช่น อายุ เพศ ประวัติการได้รับวัคซีนที่ผ่านมา ทำการวิเคราะห์ร่วมกับการรับวัคซีนตามช่วงอายุ แล้วทำการแนะนำวัคซีนที่ต้องฉีดให้กับเด็กที่มาทำการรับวัคซีน ซึ่งการที่จะสามารถวิเคราะห์เพื่อการแนะนำวัคซีนที่ถูกต้องได้ ข้อมูลที่จะนำมาวิเคราะห์นั้น ต้องมีความสมบูรณ์ ครบถ้วน และถูกต้อง ซึ่งประเด็นเหล่านี้ล้วนเป็นประเด็นสำคัญ จากการตรวจสอบเบื้องต้น พบว่า ข้อมูลที่จัดเก็บในระบบฐานข้อมูลคลินิกเด็กสุขภาพดีในปัจจุบันนั้น ยังคงมีข้อมูลไม่สมบูรณ์ และไม่ถูกต้องอยู่ โดยเฉพาะในส่วนประวัติข้อมูลคนไข้ ที่การบันทึกข้อมูลน้ำหนักและส่วนสูงของเด็กนั้น มีการบันทึกน้ำหนักและส่วนสูงที่ผิดไปจากข้อเท็จจริง ซึ่งข้อมูลที่ผิดพลาดนี้ จะส่งผลกระทบต่อการระบุวัคซีนสำหรับเด็กรายนั้น ๆ เนื่องจากการให้วัคซีนขึ้นอยู่กับช่วงอายุและพัฒนาการทางร่างกายของเด็ก

ในการเลือกวัคซีนเองก็มีปัญหา ในกรณีที่เด็กมารับวัคซีนไม่เป็นไปตามระยะเวลา หรือไม่มีประวัติของเด็กกับทางโรงพยาบาล ซึ่งทำให้แพทย์ต้องซักประวัติและตรวจสอบเกณฑ์การฉีดวัคซีน เพื่อทำการวิเคราะห์เลือกวัคซีนใหม่ทดแทนมาใช้กับเด็กรายนั้นๆ ซึ่งกระบวนการเหล่านี้ต้องใช้แพทย์ที่มีความรู้เฉพาะทาง สำหรับกรณีที่มีคนไข้จำนวนมาก อาจเกิดการล่าช้าได้

1.2 ประเด็นปัญหาของงานวิจัย

ในงานวิจัยนี้ ประกอบไปด้วยประเด็นวิจัยสำคัญ 2 ประเด็น ดังต่อไปนี้

ประเด็นที่ 1 การจัดการข้อมูล ข้อมูลที่มีอยู่ในฐานข้อมูลคลินิกเด็ก ในส่วนของการระบุชื่อวัคซีน และประวัติการรักษา บางส่วนของประวัติคนไข้ มีความไม่สมบูรณ์ ไม่ครบถ้วน หรืออาจไม่ถูกต้องอยู่ ประกอบกับข้อมูลมีปริมาณเพิ่มมากขึ้นอย่างต่อเนื่อง ผู้วิจัยจึงนำเสนอการใช้โมเดลทางคณิตศาสตร์และขั้นตอนวิธีทาง Machine Learning เข้ามาช่วยจัดการกับข้อมูล โดยเริ่มจากการทำความสะอาดข้อมูล (Data Cleansing) การเติมข้อมูล (Filling Missing Value) และการแปลงข้อมูล (Data Transformation) เป็นต้น เพื่อให้ได้

ข้อมูลที่มีประสิทธิภาพ ก่อนที่จะสามารถนำไปใช้ในขั้นตอนการแนะนำวัคซีน ในขั้นตอนถัดไป

ประเด็นที่ 2 การเลือกวัคซีน การเลือกวัคซีนสำหรับเด็กแต่ละราย อาจมีความซับซ้อน เนื่องจากการมารับวัคซีนไม่เป็นไปตามระยะเวลาที่กำหนด หรือไม่มีประวัติกับโรงพยาบาล จึงทำให้บุคลากรทางการแพทย์ต้องพิจารณาเลือกใช้วัคซีนทดแทน ให้เหมาะสมกับเด็กรายคน ในประเด็นนี้ ผู้วิจัยจึงทำการศึกษาขั้นตอนวิธีทาง Machine Learning ที่เหมาะสมกับลักษณะข้อมูลในฐานข้อมูลเด็กสุขภาพดี ในการระบุวัคซีนสำหรับเด็กแต่ละราย เพื่อระบุชนิดของวัคซีนที่สอดคล้องกับข้อมูลเด็กเฉพาะราย ที่ต้องสอดคล้อง และเหมาะสมกับช่วงอายุ และพัฒนาการางกายของเด็กแต่ละคน เพื่อให้ระบบสามารถแนะนำวัคซีนสำหรับประกอบการวินิจฉัยของบุคลากรทางการแพทย์ สำหรับการบริหารจัดการคลังวัคซีน

1.3 วัตถุประสงค์ของโครงงาน

- เพื่อพัฒนาขั้นตอนวิธีสำหรับการตรวจจับข้อมูลผิดปกติ โดยใช้วิธีการ Machine Learning ที่เหมาะสมกับลักษณะข้อมูลในฐานข้อมูลคลินิกเด็ก เพื่อเตรียมข้อมูลสำหรับนำไปใช้ในระบบแนะนำวัคซีน
- เพื่อศึกษาขั้นตอนวิธี ในการระบุชนิดของวัคซีนที่สอดคล้องกับข้อมูลรายคน สำหรับการแนะนำวัคซีนที่เหมาะสมกับช่วงอายุ พัฒนาการทางร่างกายของเด็กแต่ละคน เพื่อสนับสนุนการวินิจฉัยของแพทย์

1.4 ประโยชน์ที่คาดว่าจะได้รับ

- ได้ขั้นตอนวิธีสำหรับการตรวจจับข้อมูลผิดปกติ โดยใช้วิธีการ Machine Learning ที่เหมาะสมกับลักษณะข้อมูลในฐานข้อมูลคลินิกเด็ก เพื่อเตรียมข้อมูลสำหรับนำไปใช้ในระบบแนะนำวัคซีน
- ได้ขั้นตอนวิธี ในการระบุชนิดของวัคซีนที่สอดคล้องกับข้อมูลรายคน สำหรับการแนะนำวัคซีนที่เหมาะสมกับช่วงอายุ พัฒนาการทางร่างกายของเด็กแต่ละคน เพื่อสนับสนุนการวินิจฉัยของแพทย์

1.5 ขอบเขตของงานวิจัย

วิธีการที่นำเสนอในงานวิจัย ใช้ได้โดยตรงกับข้อมูลในฐานข้อมูลคลินิกเด็กสุขภาพดี ในส่วนของข้อมูลการให้วัคซีนเท่านั้น โดยมีขอบเขตของข้อมูลที่นำมาใช้ในงานวิจัยนี้ ดังนี้

- ฐานข้อมูลคลินิกเด็กสุขภาพดี เป็นข้อมูลคนไข้เด็กช่วงอายุ 0 ถึง 6 ปี ระหว่างวันที่ 1 มกราคม พ.ศ.2562 ถึง 31 สิงหาคม พ.ศ.2563 จากโรงพยาบาลมหาวิทยาลัยบูรพา มีข้อมูลทั้งหมด 2533 แถว มีข้อมูลผิดปกติ 4%
- ข้อมูลอื่นที่นำมาทดลองใน Phase I เพื่อเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอประกอบด้วย Stamps มีข้อมูลทั้งหมด 340 แถว มีข้อมูลผิดปกติ 9.10%, Arrhythmia มีข้อมูลทั้งหมด 450 แถว มีข้อมูลผิดปกติ 15.8%, Pima มีข้อมูลทั้งหมด 768 แถว มีข้อมูลผิดปกติ 34.9% และ Parkinson มีข้อมูลทั้งหมด 195 แถว มีข้อมูลผิดปกติ 75.4%
- ข้อมูลอื่นที่นำมาทดลองใน Phase II เพื่อเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอประกอบด้วย Wine มีข้อมูลทั้งหมด 1599 แถว มีมิติข้อมูล 12 มิติ และมีคลาส 6 คลาส และ Iris มีข้อมูลทั้งหมด 150 แถว มีมิติข้อมูล 6 มิติ และมีคลาส 3 คลาส

1.6 แผนการดำเนินงานวิจัย

ขั้นตอนการดำเนินงาน	ระยะเวลาในการดำเนินงาน																							
	พ.ศ.2562				พ.ศ.2563												พ.ศ.2564							
	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8
1.ศึกษาวรรณกรรมที่เกี่ยวข้อง																								
2.ทดลองอัลกอริทึมของงานวิจัยในอดีตและสรุป																								
3.รวบรวมและวิเคราะห์																								
4.เขียนรูปเล่ม																								
5.นำเสนองานวิจัย																								

ตาราง 1-1 แผนการดำเนินงานวิจัย

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง ได้แก่ การให้วัคซีนของเด็ก การเรียนรู้ของเครื่อง (Machine Learning) ขั้นตอนวิธี K- Nearest Neighbors ระบบแนะนำ (Recommender System) และงานวิจัยที่เกี่ยวข้อง ดังนี้

2.1 **การให้วัคซีนเด็ก**

วัคซีน หมายถึง ชีววัตถุหรือแอนติเจนที่ถูกผลิตมาจากเชื้อโรค หรือพิษของเชื้อโรคที่ถูกทำให้ไม่สามารถก่อโรคได้ แต่ยังคงกระตุ้นให้ร่างกายสร้างแอนติบอดี หรือภูมิคุ้มกันได้ เมื่อร่างกายได้รับวัคซีนครั้งแรก ร่างกายจะใช้เวลาระยะหนึ่งสร้างแอนติบอดีขึ้นมาพร้อมทั้งจดจำแอนติเจนไว้ เมื่อเวลาผ่านไประยะหนึ่งปริมาณแอนติบอดีนี้จะค่อยๆลดลง แต่หากร่างกายได้รับวัคซีนนั้นอีกครั้งที่สอง ร่างกายจะตอบสนองด้วยการสร้างแอนติบอดีได้เร็วขึ้น และปริมาณมากขึ้นกว่าครั้งแรก จึงเป็นเหตุผลว่าทำไมจึงต้องมีการกระตุ้นด้วยการให้วัคซีนครั้งที่ 2 หรือ 3 ในทางตรงกันข้ามการได้รับวัคซีนครั้งที่ 2 เร็วกว่ากำหนดอาจไม่กระตุ้นให้ร่างกายสร้างแอนติบอดีเพิ่มขึ้น เนื่องจากแอนติบอดีที่เกิดขึ้นในครั้งแรกยังมีระดับสูงอยู่ ทำให้แอนติเจนที่เข้าสู่ร่างกายครั้งนี้ไปทำปฏิกิริยากับแอนติบอดีที่อยู่ในร่างกายหมด ไม่มีเหลือไปกระตุ้นให้ร่างกายสร้างแอนติบอดีอีก

การให้วัคซีนเป็นวิธีการหนึ่งที่ดีในการป้องกันโรคติดเชื้อที่มีประสิทธิภาพและเป็นที่ยอมรับทั่วโลก หลักการทำงานของวัคซีนคร่าวๆ คือการนำเชื้อโรคที่ตายหรือทำให้อ่อนแอเข้าสู่ร่างกายด้วยวิธีการต่างๆ เช่น การฉีด เพื่อให้เชื้อโรคหรือสิ่งแปลกปลอม (แอนติเจน) ไปกระตุ้นระบบภูมิคุ้มกันในร่างกาย โดยเซลล์เม็ดเลือดขาวชนิดลิมโฟไซต์จะเป็นตัวกำจัดเชื้อ พร้อมกับจดจำลักษณะของแอนติเจนชนิดนี้ เพื่อสามารถสร้างแอนติบอดีให้เข้ากำจัดเชื้อได้อย่างรวดเร็วในครั้งต่อไปที่มีแอนติเจนชนิดนี้เข้าสู่ร่างกาย การฉีดวัคซีนจึงเป็นเหมือนการเข้าไปกระตุ้นให้ร่างกายรู้จักเชื้อโรคและสร้างแอนติบอดีที่สามารถรับมือกับเชื้อโรคนั้น ๆ ได้ทันท่วงที

ในปัจจุบันการให้วัคซีนเริ่มให้ตั้งแต่เด็ก เนื่องจากเด็กมีความเสี่ยงที่จะเกิดโรคได้ เพราะร่างกายของเด็กมีระดับภูมิคุ้มกันที่ต่ำ ถ้าหากไม่ได้รับการฉีดวัคซีนเพื่อป้องกัน เมื่อเด็กเกิดการติดเชื้อ นอกจากจะทำให้การรักษาลำบาก เสียเวลา อีกทั้งต้องเสียค่าใช้จ่ายที่มากขึ้น ดังนั้นการให้

วัคซีนจึงมีความสำคัญต่อสุขภาพอนามัย เพื่อป้องกันการติดเชื้อ การแพร่ระบาดในชุมชน จะทำให้เด็กแข็งแรง มีพัฒนาการที่ดี ลดความเสี่ยงในการเกิดโรค

สำหรับการให้วัคซีนในเด็กนั้นจะมี 2 กลุ่มใหญ่ คือ วัคซีนพื้นฐานเป็นวัคซีนจำเป็นที่กระทรวงสาธารณสุขกำหนดให้เด็กทุกคนจะต้องได้รับ เช่น วัคซีนป้องกันโรคโปลิโอ วัคซีนป้องกันโรคหัด วัคซีนป้องกันโรคโรคไข้สมองอักเสบเจอี วัคซีนป้องกันโรคตับอักเสบบี เป็นต้น ส่วนอีกกลุ่ม คือ วัคซีนเสริมหรือวัคซีนทดแทน เป็นวัคซีนที่มีประโยชน์ในการป้องกันโรคเพิ่มเติมจากวัคซีนพื้นฐาน แต่ยังไม่มีความสำคัญด้านสาธารณสุขในลำดับต้นๆ เช่น วัคซีนป้องกันโรคตับอักเสบเอ วัคซีนป้องกันโรคโรต้า วัคซีนป้องกันโรคอีสุกอีใส วัคซีนป้องกันโรคนิวโมคอคคัสชนิดคอนจูเกต เป็นต้น ซึ่งการให้วัคซีนในเด็กแต่ละชนิดนั้นไม่สามารถให้พร้อมกันทีเดียวได้ แต่จะให้ตามช่วงอายุของเด็กซึ่งหลักเกณฑ์การให้วัคซีนก็จะแตกต่างกัน เนื่องจากสภาพร่างกายของเด็กแต่คนไม่เหมือนกัน บางคนอาจมีอาการแพ้ เด็กบางคนจำเป็นต้องได้รับวัคซีนเสริม แต่บางคนกลับไม่จำเป็น วัคซีนบางชนิดต้องฉีดกระตุ้น 2-3 เข็ม ในระยะเวลาที่ห่างกัน ซึ่งเด็กบางคนอาจจะมารับวัคซีนตรงหรือไม่ตรง ปัญหาเหล่านี้แพทย์ผู้ที่ให้วัคซีนต้องพิจารณา คำนึงถึงสุขภาพเด็กแต่ละคนว่า ต้องได้รับวัคซีนชนิดไหน จากบรรดาวัคซีนหลากหลายชนิด หลายผู้ผลิต เพื่อให้เกิดความเหมาะสมกับเด็กแต่ละคน

การฉีดวัคซีน คือ การสร้างภูมิคุ้มกัน เพื่อป้องกันการติดต่อของโรคติดต่อร้ายแรง เชื้อวัคซีนผลิตจากเชื้อไวรัส หรือ เชื้อแบคทีเรียที่อ่อนตัวแล้ว หรือส่วนประกอบอื่นของเชื้อ ตัวของวัคซีนเองก่อให้เกิดอาการน้อยมาก แต่หากจะทำการสร้างภูมิต้านทานให้ร่างกาย ซึ่งจะป้องกันหรือช่วยลดความเสี่ยงของเด็กที่อาจป่วยเป็นโรคที่ทำการฉีดวัคซีนนั้นๆ

วัคซีนแบ่งออกเป็น 3 ชนิด มีดังนี้

- **วัคซีนเชื้อตาย (Killed Vaccine)** หมายถึง วัคซีนที่ผลิตขึ้นโดยใช้เชื้อโรคทั้งตัวที่ตายแล้ว ตัวอย่างวัคซีนในกลุ่มนี้ ได้แก่ วัคซีนไอกรนชนิดทั้งเซลล์ วัคซีนตับอักเสบเอ วัคซีนอหิวาตกโรคชนิดฉีด วัคซีนพิษสุนัขบ้า และวัคซีนไข้สมองอักเสบเจอีชนิดเชื้อตาย วัคซีนโปลิโอชนิดฉีด และผลิตจากส่วนประกอบบางส่วนของเชื้อโรค ตัวอย่างวัคซีนในกลุ่มนี้ ได้แก่ วัคซีนตับอักเสบบี วัคซีนไข้หวัดใหญ่ วัคซีนฮิบ วัคซีนไอกรนชนิดไร้เซลล์ และวัคซีนนิวโมคอคคัส หรือผลิตจากโปรตีนส่วนประกอบของเชื้อที่ผลิตมาใหม่โดยอาศัยหลักวิศวพันธุศาสตร์
- **วัคซีนเชื้อเป็นอ่อนฤทธิ์ (Live Attenuated Vaccine)** หมายถึง วัคซีนที่ผลิตขึ้นโดยใช้เชื้อโรคมาทำให้อ่อนฤทธิ์ลงจนไม่สามารถทำให้เกิดโรคแต่เพียงพอที่จะกระตุ้นภูมิคุ้มกันของร่างกายได้ วัคซีนในกลุ่มนี้ได้แก่ วัคซีนโปลิโอชนิดกิน วัคซีน

รวมหัด หัดเยอรมัน คางทูม วัคซีนอีสุกอีใส วัคซีนวัณโรค วัคซีนโรต้า วัคซีนไข้หวัดใหญ่ชนิดพ่นจมูก

- **วัคซีนประเภทท็อกซอยด์ (Toxoid)** หมายถึง วัคซีนที่ผลิตโดยการนำพิษของจุลชีพที่เป็นส่วนสำคัญในการก่อโรคมาทำให้หมดฤทธิ์แต่ยังสามารถกระตุ้นให้เกิดภูมิคุ้มกันโรคได้ เช่นวัคซีนคอตีบ และวัคซีนบาดทะยัก

สำหรับการให้วัคซีนในเด็กนั้นจะมี 2 กลุ่มใหญ่ คือ วัคซีนพื้นฐานเป็นวัคซีนจำเป็นที่กระทรวงสาธารณสุขกำหนดให้เด็กทุกคนจะต้องได้รับ เช่น วัคซีนป้องกันโรคโปลิโอ วัคซีนป้องกันโรคหัด วัคซีนป้องกันโรคโรคไข้สมองอักเสบเจอี วัคซีนป้องกันโรคตับอักเสบบี เป็นต้น ส่วนอีกกลุ่ม คือ วัคซีนเสริมหรือวัคซีนทดแทน เป็นวัคซีนที่มีประโยชน์ในการป้องกันโรคเพิ่มเติมจากวัคซีนพื้นฐาน แต่ยังไม่มีความสำคัญด้านสาธารณสุขในลำดับต้นๆ เช่น วัคซีนป้องกันโรคตับอักเสบเอ วัคซีนป้องกันโรคโรต้า วัคซีนป้องกันโรคอีสุกอีใส วัคซีนป้องกันโรคนิวโมคอคคัสชนิดคอนจูเกต เป็นต้น

รายละเอียดของวัคซีน มีดังนี้

1. **วัคซีนบีซีจี**
    1.1 ฉีด 0.1 มล. ในชั้นผิวหนังที่ไหล่ซ้าย ไม่ควรฉีดที่สะโพก
    1.2 ถ้าไม่มีแผลเป็นเกิดขึ้น และไม่มีหลักฐานว่าเคยได้รับวัคซีนบีซีจีมาก่อน ให้ได้ฉีดทันที
    1.3 ถ้าเคยได้รับวัคซีนบีซีจีมาก่อน ไม่ต้องฉีดซ้ำแม้ไม่มีแผลเป็น
2. **วัคซีนตับอักเสบบี**
    2.1 เด็กทุกคนต้องได้รับอย่างน้อย 3 ครั้ง ถ้าไม่มีข้อห้าม และเข็มสุดท้ายต้องอายุมากกว่าหรือเท่ากับ 6 เดือน
    2.2 ทารกคลอดจากมารดาที่มี HBsAg เป็นลบ ให้วัคซีนจำนวน 3 ครั้งเมื่อแรกเกิด อายุ 1-2 เดือน และอายุ 6 เดือนตามลำดับ กรณีไม่ทราบผลเลือดมารดา ควรให้วัคซีนครั้งที่ 1 ภายใน 12 ชม. หลังคลอด ครั้งที่ 2 และ 3 ที่อายุ 1 เดือน และ 6 เดือนตามลำดับ
    2.3 ทารกที่คลอดจากมารดาที่มี HBsAg เป็นบวก (โดยเฉพาะถ้า HBeAg เป็นบวกด้วย) พิจารณาให้ HBIG 0.5 มล. ภายใน 12 ชม. หลังคลอด และให้วัคซีนครั้งที่ 1 พร้อมๆกันคนละตำแหน่งกับ HBIG
        - กรณีทารกได้รับ HBIG ให้ฉีดวัคซีนครั้งที่ 2 เมื่ออายุ 1-2 เดือน และครั้งที่ 3 เมื่ออายุ 6 เดือน

- กรณีทารกไม่ได้รับ HBIG ควรให้วัคซีนครั้งที่ 1 ภายใน 12 ชม. หลังคลอด ครั้งที่ 2 เมื่ออายุ 1 เดือน และครั้งที่ 3 เมื่ออายุ 6 เดือน

2.4 ในกรณีที่มาทราบภายหลังว่ามารดามี HBsAg เป็นบวก ควรพิจารณาให้ HBIG ถ้าทารกได้รับวัคซีนมาแล้วไม่เกิน 7 วัน

2.5 ตามแผนการสร้างเสริมภูมิคุ้มกันโรคของกระทรวงสาธารณสุข ใช้วัคซีนรวมที่มี คอตีบ-บาดทะยัก-ไอกรน-ตับอักเสบบี(DTP-HB) ที่อายุ 2,4 และ 6 เดือน แต่ถ้ามารดามี HBsAg เป็นบวก และทารกไม่ได้ HBIG ควรให้วัคซีนตับอักเสบบีแบบเดี่ยวเพิ่มตอนอายุ 1 เดือนด้วย (รวมเป็น 5 ครั้ง)

2.6 เด็กที่ไม่เคยได้รับวัคซีนตับอักเสบบีมาก่อน เด็กอายุน้อยกว่า 11 ปี สามารถฉีดวัคซีนได้ในเดือนที่ 0, 1, 6 ตามลำดับ ส่วนเด็กอายุ 11-15 ปี อาจใช้วัคซีน HBVax Pro™ (ผลิตโดย MSD) ฉีดเพียง 2 ครั้ง ในเดือนที่ 0 และเดือนที่ 4-6 โดยใช้วัคซีนขนาด 1.0 มล. เท่าผู้ใหญ่

2.7 เด็กที่คลอดจากมารดาที่มี HBsAg เป็นบวก พิจารณาตรวจ HBsAg และ anti-HBs เมื่ออายุประมาณ 9-12 เดือน

3. **วัคซีนคอตีบ-บาดทะยัก-ไอกรน**

3.1 สามารถใช้ชนิดไร้เซลล์ (DTaP) แทนชนิดทั้งเซลล์ (DTwP) ได้ทุกครั้ง

3.2 หากใช้ DTaP ควรใช้ชนิดเดียวกันทั้งสามครั้งเมื่ออายุ 2, 4, 6 เดือน หากไม่สามารถหาชนิดเดียวกันได้ ให้ใช้ชนิดใดแทนก็ได้

3.3 สำหรับเข็มกระตุ้นที่ 18 เดือน อาจใช้ DTwP หรือ DTaP ชนิดใดก็ได้

3.4 เมื่ออายุ 4-6 ปี อาจใช้ DTwP, DTaP หรือ Tdap (Boostrix™ หรือ Adacel™)ก็ได้

3.5 เด็กอายุ 11-12 ปี ควรได้รับการฉีด Td หรือ Tdap หรือ TdaP (Boostagen™) ไม่ว่าจะเคยได้รับ Tdap เมื่ออายุ 4-6 ปี มาก่อนหรือไม่ หลังจากนั้นควรฉีดกระตุ้นด้วย Td ทุก 10 ปี

3.6 ผู้ควรได้รับ Tdap หรือ TdaP 1 ครั้ง ไม่ว่าจะเคยได้ TT หรือ Td มานานเท่าใดก็ตาม จากนั้นให้ฉีดกระตุ้นด้วย Td ทุก 10 ปี

3.7 หญิงตั้งครรภ์ควรได้รับ Tdap 1 ครั้งที่อายุครรภ์ 27-36 สัปดาห์ทุกการตั้งครรภ์

4. **วัคซีนโปลิโอ**

4.1 ให้หยอด bivalent OPV (type1, 3) 5 ครั้ง ร่วมกับฉีด IPV 1 ครั้งที่อายุ 4 เดือน

4.2 สามารถใช้ชนิดฉีดแทนชนิดกินได้ทุกครั้ง หากใช้ชนิดฉีดอย่างเดียวโดยตลอดอาจให้เพียง 4 ครั้ง โดยงดเมื่ออายุ 18 เดือนได้

5. **วัคซีนหัด-หัดเยอรมัน-คางทูม**

5.1 ให้วัคซีนครั้งแรกเมื่ออายุ 9-12 เดือน และครั้งที่ 2 เมื่ออายุ 2½ ปี ในพื้นที่ที่มีรายงานโรคหัดจำนวนน้อย อาจฉีดเข็มแรกหลังอายุ 12 เดือน และครั้งที่ 2 ที่อายุ 2½-4 ปี (แนะนำให้ที่อายุ 2½ ปี ตามนโยบายของกระทรวงสาธารณสุข)

5.2 ในกรณีที่มีการระบาดหรือสัมผัสโรค อาจฉีดเข็มแรกได้ตั้งแต่อายุ 6 เดือน เป็นต้นไป ในกรณีที่เข็มแรกได้รับก่อนอายุ 9 เดือนให้ฉีดเข็มที่ 2 ที่อายุ 12 เดือน และเข็มที่ 3 ที่อายุ 2½-4 ปี

5.3 ในกรณีที่ฉีดเข็มที่ 1 แล้ว และเกิดการระบาดขึ้น เข็มที่ 2 สามารถให้ก่อนอายุ 2½ ปีได้ แต่ต้องห่างจากเข็มแรกอย่างน้อย 3 เดือน

5.4 ในกรณีที่ต้องการฉีดวัคซีน หัด-หัดเยอรมัน-คางทูม และอีสุกอีใสในเวลาเดียวกัน สามารถใช้วัคซีนรวม หัด-หัดเยอรมัน-คางทูม-อีสุกอีใส (MMRV) แทนการฉีดแบบแยกเข็มได้ทุกครั้งในเด็กอายุตั้งแต่ 1-12 ปี การใช้วัคซีนรวม MMRV ที่อายุ 2½-4 ปีแทนการฉีดวัคซีนแบบแยกเข็มพบมีอาการข้างเคียงไม่แตกต่างกัน

5.5 การฉีดวัคซีนรวม MMRV ในเด็กอายุ 12-23 เดือนโด๊สแรกมีโอกาสเกิดการชักจากไข้ได้มากกว่าการฉีดแยกเข็ม สำหรับกรณีที่เคยได้วัคซีน MMR หรือ VZV มาก่อน แนะนำให้วัคซีนรวม MMRV ห่างจากวัคซีน MMR และ VZV ครั้งก่อน อย่างน้อย 3 เดือน

6. **วัคซีนไข้สมองอักเสบเจอี**

6.1 วัคซีนชนิดเชื้อไม่มีชีวิต (inactivated vaccine) ปัจจุบันมีชนิดทำจากสายพันธุ์ P3 เพาะเลี้ยงใน vero cell (JEVAC™) ฉีด 3 ครั้ง เริ่มเมื่ออายุ 6 เดือนขึ้นไป เข็มต่อมาอีก 4 สัปดาห์ และ 1 ปีตามลำดับ

6.2 วัคซีนชนิดเชื้อมีชีวิต (live JE; สายพันธุ์ SA 14-14-2) เริ่มฉีดที่อายุ 9-12 เดือน ให้ฉีด 2 ครั้ง มี 2 ชนิด คือ CD-JEVAX™ ฉีดเข็มที่ 2 อีก 3-12 เดือนต่อมา อีกชนิดคือ Chimeric JE (IMOJEV™/THAI JEV™) ฉีดเข็มที่ 2 อีก 12-24 เดือนต่อมา live JE ทั้งสองชนิดสามารถใช้แทนกันได้ การฉีดตามแผนฯของกระทรวงสาธารณสุขทั้งสองชนิดให้ฉีดห่างกัน 12-18 เดือน

6.3 สามารถใช้วัคซีนชนิด live JE แทนชนิด inactivated JE ได้ ทั้งในการฉีดชุดแรกและการฉีดกระตุ้น ในกรณีที่เคยได้รับ inactivated JE มาก่อน และต้องการฉีดต่อด้วย live JE vaccine ให้พิจารณาฉีดตามตาราง

ประวัติการฉีดวัคซีน inactivated JE ในอดีต	ข้อแนะนำในการฉีดวัคซีน live-attenuated JE
1 เข็ม	ฉีด 2 เข็ม ห่างกัน 3-24 เดือน(แล้วแต่ชนิดของวัคซีน)
2-3 เข็ม	ฉีด 1 เข็ม ห่างจากเข็มสุดท้าย 1 ปี
≥4 เข็ม	ไม่จำเป็นต้องฉีดอีก*

ตาราง 2-1 พิจารณาการฉีด live JE

*****อาจพิจารณาฉีด live JE 1 เข็ม ห่างจากเข็มสุดท้าย 1

6.4 กรณีที่เคยได้รับ live JE มาก่อน 1 ครั้ง หากจำเป็นต้องฉีดกระตุ้นด้วย inactivated JE ให้ฉีดอีก 1 ครั้ง ห่างกันอย่างน้อย 12 เดือน

7. **วัคซีนไข้หวัดใหญ่**

7.1 พิจารณาให้ฉีดในเด็กอายุ 6 เดือนขึ้นไปถึง 18 ปี (ตามเอกสารกำกับยา) โดยเฉพาะเด็กอายุน้อยกว่า 2 ปี และเด็กที่มีความเสี่ยงที่จะเป็นโรครุนแรง เช่น เด็กที่เป็นโรคปอดเรื้อรัง (รวมหอบหืด) โรคหัวใจ โรคอ้วนที่มี BMI > 35 ภูมิคุ้มกันบกพร่อง หญิงตั้งครรภ์และโรคเรื้อรังอื่นๆ เป็นต้น โดยในกลุ่มเหล่านี้มีวัคซีนจัดสรรให้ปีละครั้งตามแผนของกระทรวงฯ แนะนำให้ฉีดก่อนเข้าฤดูฝน อย่างไรก็ตามสามารถฉีดได้ตลอดปี

7.2 ถ้าอายุน้อยกว่า 9 ปี การฉีดในครั้งแรกต้องฉีดสองเข็มห่างกัน 1 เดือน กรณีที่ปีแรกได้ฉีดไปเพียงครั้งเดียว ปีถัดมาให้ฉีดสองครั้ง จากนั้นจึงสามารถฉีดปีละครั้งได้

7.3 ในเด็กอายุน้อยกว่า 3 ปี ให้ได้ทั้งขนาด 0.25 หรือ 0.5 มล. ยกเว้น Fluarix Tetra ที่ให้ฉีด 0.5 มล.เท่านั้น(ตามเอกสารกำกับยา)

7.4 สามารถใช้วัคซีนไข้หวัดใหญ่ชนิด 3 หรือ 4 สายพันธุ์ ทดแทนกันได้

8. **วัคซีนเอชพีวี**

8.1 มี 2 ชนิดคือ ชนิด 2 สายพันธุ์ (bivalent มีสายพันธุ์ 16, 18)และชนิด 4 สายพันธุ์ (quadrivalent มีสายพันธุ์ 6, 11, 16, 18) หากต้องการให้ป้องกันหูดหงอนไก่ด้วยต้องใช้วัคซีนชนิด 4 สายพันธุ์

8.2 แนะนำให้ฉีดในหญิงและชาย อายุ 9-26 ปี (เน้นให้ฉีดในช่วงอายุ 11-12 ปี) โดยฉีด 3 เข็ม ในเดือนที่ 0, 1-2, และ 6

8.3 ในวัยรุ่นที่แข็งแรงดี หากฉีดเข็มแรกก่อนอายุ 15 ปี ให้ฉีด 2 เข็มได้ ที่ 0, 12 เดือน

8.4 ตามแผนฯ ของกระทรวงสาธารณสุขให้ฉีดในเด็กหญิงชั้นประถมปีที่ 5 ให้ 2 เข็ม ห่างกัน 6 เดือน

8.5 ประสิทธิภาพของวัคซีนจะสูงหากฉีดในผู้ที่ไม่เคยมีเพศสัมพันธ์มาก่อน

8.6 การฉีดในผู้ที่มีอายุมากกว่า 26 ปี พิจารณาให้ได้เป็นกรณีๆไป อายุที่แนะนำเป็นไปตามเอกสารกำกับยา

9. **วัคซีนฮิบ**

9.1 ปัจจุบันมีชนิด conjugate กับ PRP-T ในเด็กแนะนำให้ 3 ครั้งเมื่ออายุ 2, 4 และ 6 เดือน

9.2 การฉีดเข็มกระตุ้นที่อายุ 12-18 เดือน อาจไม่จำเป็นต้องฉีดในเด็กแข็งแรง ควรฉีดในผู้ที่มีความเสี่ยง

9.3 ไม่จำเป็นต้องฉีดวัคซีนฮิบในเด็กปกติที่อายุ 2 ปีขึ้นไป

9.4 หากเริ่มฉีดช้า ให้พิจารณาฉีดตามตาราง

อายุที่เริ่มฉีด	**เดือนที่ของการฉีด PRP-T**
2-6 เดือน	0, 2, 4, ฉีดกระตุ้นอายุ 12-18 เดือน
7-11 เดือน	0, 2, ฉีดกระตุ้นอายุ 12-18 เดือน
12-24 เดือน	เข็มเดียว
>24 เดือน เฉพาะผู้ที่เสี่ยง*	0, 2

ตาราง 2-2 การฉีดวัคซีน Hib

*****ผู้ที่เสี่ยงต่อโรคฮิบ เช่น ผู้ที่ภูมิคุ้มกันบกพร่อง ไม่มีม้าม หรือม้ามทำงานผิดปกติ

10. **วัคซีนตับอักเสบเอ**

10.1 วัคซีนชนิดเชื้อไม่มีชีวิต (inactivated vaccine) ฉีดได้ตั้งแต่อายุ 1 ปีขึ้นไป โดยฉีด 2 เข็มห่างกัน 6-12 เดือน อาจใช้ต่างชนิดได้ในการฉีดแต่ละครั้ง

10.2 วัคซีนชนิดเชื้อมีชีวิต (live vaccine) ฉีดได้ตั้งแต่อายุ 18 เดือนขึ้นไป เพียงเข็มเดียว

11. **วัคซีนอีสุกอีใส**

11.1 ฉีดได้ตั้งแต่อายุ 1 ปีขึ้นไป แนะนำให้ฉีดเข็มแรกอายุ 12-18 เดือน

11.2 อาจพิจารณาให้ฉีดเข็มที่ 2 เมื่ออายุ 2½-4 ปี อาจฉีดเข็มที่ 2 ก่อนอายุ 4 ปี ได้ในกรณีที่มีการระบาด โดยต้องห่างจากเข็มแรกอย่างน้อย 3 เดือน อาจใช้วัคซีน MMRV แทน MMR และ VZV แบบแยกเข็ม (ดูในข้อ 5)

11.3 ถ้าอายุมากกว่า 13 ปี ให้ฉีดสองเข็มห่างกันอย่างน้อย 1 เดือน

12. **วัคซีนนิวโมคอคคัส ชนิดคอนจูเกต**

12.1 ควรให้ในผู้ที่มีความเสี่ยงต่อโรคนี้ชนิดรุกราน (invasive disease) หรือรุนแรง (severe) ดังตาราง และในเด็กแข็งแรงปกติที่อายุน้อยกว่า 5 ปี ที่ประสงค์จะป้องกันโรค

12.2 ปัจจุบันมีวัคซีน ชนิด 10 สายพันธุ์ (PCV10) และ 13 สายพันธุ์ (PCV13) ให้ 3 ครั้ง เมื่ออายุ 2, 4 และ 6 เดือน และให้ฉีดกระตุ้นที่อายุ 12-15 เดือน โดยห่างจากเข็มสุดท้ายอย่างน้อย 2 เดือน หากเริ่มฉีดช้าให้ฉีดตามตาราง

12.3 ในเด็กปกติ อาจพิจารณาให้ฉีดแบบ 2+1 (รวมเป็นการฉีด 3 ครั้ง) คือฉีดเมื่ออายุ 2, 4, และ 12-15 เดือน

อายุที่เริ่มฉีด	จำนวนครั้งที่ฉีด	การฉีดกระตุ้น
เด็กปกติและเด็กเสี่ยง 2-6 เดือน	PCV 3 ครั้ง ห่างกัน 6-8 สัปดาห์	PCV 1 ครั้ง อายุ 12-15 เดือน
เด็กปกติและเด็กเสี่ยง 7-11 เดือน	PCV 2 ครั้ง ห่างกัน 6-8 สัปดาห์	PCV 1 ครั้ง อายุ 12-15 เดือน
เด็กปกติและเด็กเสี่ยง 12-23 เดือน	PCV 2 ครั้ง ห่างกัน 6-8 สัปดาห์	ไม่ต้องฉีด
เด็กปกติ 2-5 ปี	PCV 10 ให้ 2 ครั้ง PCV13 ให้ 1 ครั้ง	ไม่ต้องฉีด
*เด็กเสี่ยง - อายุ 2-5 ปี - อายุ 2-6 ปี - >6-18 ปี	 PCV10 ให้ 2 ครั้ง ห่างกัน 8 สัปดาห์ PCV13 ให้ 2 ครั้ง ห่างกัน 8 สัปดาห์ PCV13 ให้ 1 ครั้ง	 ฉีดกระตุ้นด้วย PS-23 1 เข็ม ห่างจาก PCV เข็มสุดท้าย 8 สัปดาห์

ตาราง 2-3 การฉีดวัคซีน PCV

หมายเหตุ: PCV =Pneumococcal conjugate vaccine, PS-23= 23-Valent pneumococcal polysaccharide vaccine

- *เด็กเสี่ยง คือเด็กที่มีโอกาสเป็นโรคติดเชื้อนิวโมคอคคัสอย่างรุนแรงมากกว่าเด็กปกติ ได้แก่ เด็กที่เป็นโรคภูมิคุ้มกันบกพร่องจากสาเหตุต่างๆ ภาวะที่ไม่มีม้าม ธาลัสซีเมีย โรคเรื้อรังของ

อวัยวะต่างๆ เช่น โรคปอด (รวมทั้งหอบหืดรุนแรง) โรคหัวใจ โรคตับ โรคไต เบาหวาน และโรคที่เสี่ยงต่อเยื่อหุ้มสมองอักเสบ เช่น CSF leak, cochlear implantation

- สำหรับเด็กที่อยู่ในสถานเลี้ยงเด็กกลางวันไม่จัดเป็นกลุ่มเสี่ยงแต่อาจพิจารณาให้วัคซีนได้
- *เด็กเสี่ยงทั้งหมด ควรได้รับ PCV ดังตาราง และเด็กเสี่ยงที่มีอายุตั้งแต่ 2 ปีขึ้นไป ควรให้ฉีดวัคซีน PS-23 ด้วยเสมอ ไม่ว่าจะสามารถฉีด PCV ได้หรือไม่ก็ตาม และหากเป็นเด็กเสี่ยงประเภทภูมิคุ้มกันบกพร่อง ภาวะไม่มีม้าม หรือธาลัสซีเมีย ควรฉีด PS-23 ซ้ำอีก 1 ครั้ง ห่างจากครั้งแรก 5 ปี การฉีด PCV ก่อน แล้วตามด้วย PS-23 จะให้ผลการสร้างภูมิคุ้มกันที่ดีกว่าการฉีด PS-23 เพียงอย่างเดียว หรือฉีด PS-23 แล้วตามด้วย PCV

13. **วัคซีนโรต้า**

13.1 ชนิด monovalent (human) ให้กิน 2 ครั้ง เมื่ออายุประมาณ 2 และ 4 เดือน

13.2 ชนิด pentavalent (bovine-human) ให้กิน 3 ครั้ง เมื่ออายุประมาณ 2, 4 และ 6 เดือน

13.3 วัคซีนทั้งสองชนิด สามารถเริ่มให้ครั้งแรกได้ เมื่ออายุ 6-15 สัปดาห์ และครั้งสุดท้ายอายุไม่เกิน 8 เดือน โดยแต่ละครั้งห่างกันไม่น้อยกว่า 4 สัปดาห์

13.4 ควรใช้วัคซีนชนิดเดียวกันจนครบ หากจำเป็นต้องใช้วัคซีนต่างชนิดกันในแต่ละครั้งหรือไม่ทราบชนิดของวัคซีนที่ได้รับในครั้งก่อน ต้องให้วัคซีนทั้งหมด 3 ครั้ง

13.5 สามารถให้วัคซีนโรต้าร่วมกับวัคซีนโปลิโอชนิดกินได้

13.6 ห้ามใช้วัคซีนนี้ในผู้ที่มีภูมิคุ้มกันบกพร่อง severe combined immune deficiency (SCID) และในเด็กที่มีประวัติลำไส้กลืนกัน

14. **วัคซีนไข้เลือดออก**

14.1 วัคซีนเป็นชนิด chimeric ของ yellow fever virus สายพันธุ์ 17D กับไข้เลือดออก 4 สายพันธุ์ (DEN 1-4)

14.2 ให้ฉีดในกลุ่มอายุ 9-45 ปี ฉีด 3 เข็ม เดือนที่ 0, 6 และ 12 ในผู้ที่เคยมีการติดเชื้อมาก่อน

14.3 ไม่แนะนำให้ฉีดในผู้ที่ไม่เคยมีการติดเชื้อมาก่อน

14.4 ในกรณีที่ไม่ทราบว่าติดเชื้อมาก่อนหรือไม่ ก่อนที่จะให้วัคซีนต้องอธิบายให้ทราบความเสี่ยงที่อาจเกิดไข้เลือดออกที่ต้องนอนโรงพยาบาลมากขึ้นได้หลังฉีดแล้ว 2 ปีเป็นต้นไปหากเป็นผู้ที่ไม่เคยติดเชื้อมาก่อน

การเว้นระยะห่างของวัคซีน

อายุที่แนะนำให้วัคซีน อายุน้อยที่สุดที่สามารถให้วัคซีนได้ และระยะห่างแต่ละโด๊ส ดังแสดงตารางที่ ไม่ควรให้วัคซีนอายุน้อยกว่าที่แนะนำ และเว้นระยะห่างสั้นกว่าที่แนะนำ ยกเว้นบางกรณีที่ต้องการให้มีภูมิคุ้มกันเร็ว เช่น ต้องเดินทางไปในพื้นที่ที่มีโรคชุกชุม หรือกรณีที่มีการระบาด เช่น โรคหัด สามารถให้วัคซีนในเด็กอายุน้อยกว่า 6 เดือนได้ แต่ไม่นับรวมอยู่ในโปรแกรมการให้วัคซีนปกติ

การให้วัคซีนอายุน้อยกว่าที่แนะนำ และเว้นระยะห่างสั้นกว่าที่แนะนำจะมีผลต่อการตอบสนองการสร้างภูมิคุ้มกันที่อาจไม่เพียงพอ อย่างไรก็ตามการให้เร็วกว่าอายุน้อยที่สุด 2-3 วันไม่น่าจะมีผลต่อการสร้างภูมิคุ้มกัน จึงพอยอมรับหากให้วัคซีนเร็วกว่าอายุหรือระยะห่างน้อยที่สุด<4 วัน ยกเว้นตารางการให้วัคซีนป้องกันโรคพิษสุนัขบ้าที่มีตารางเฉพาะ ถ้าได้วัคซีนเร็วกว่าอายุหรือระยะห่างน้อยที่สุด> 5 วัน ควรให้วัคซีนใหม่ โดยโด๊สต่อไปให้นับระยะห่างหลังจากโด๊สล่าสุดซึ่งเป็นโด๊สไม่ถูกต้อง ตัวอย่างเช่น วัคซีนฮิบควรให้ห่างกันอย่างน้อย 4 สัปดาห์ ถ้าได้วัคซีนโด๊สที่สองห่างจากโด๊สแรก 2 สัปดาห์ ให้ถือว่าโด๊สที่สองให้ไม่ถูกต้อง ต้องให้ใหม่โดยห่างจากโด๊สสุดท้าย (ซึ่งในที่นี้คือโด๊สที่สอง) 4 สัปดาห์ และจะนับวัคซีนที่ให้ซ้ำเป็นโด๊สที่สองแทน กรณีถ้าเป็นวัคซีนเชื้อเป็น อย่าลืมต้องเว้นระยะห่างแต่ละโด๊สอย่างน้อย 28 วัน เช่น ถ้าให้วัคซีนอีสุกอีใสโด๊สแรกเร็วกว่าอายุ 1 ปี เช่น 11 เดือน 2 สัปดาห์ ต้องให้วัคซีนใหม่ ที่อายุมากกว่า 1 ปี และต้องห่างจากโด๊สเดิม 4 สัปดาห์

วัคซีนเชื้อเป็นสามารถให้พร้อมกันหลายชนิดในวันเดียวกัน แต่หากจะให้ไม่พร้อมกัน ควรเว้นระยะเวลาให้ห่างกันอย่างน้อย 1 เดือน สำหรับวัคซีนเชื้อตายจะให้ห่างกันนานเท่าใดก็ได้

วัคซีนและโด๊สที่ให้ (vaccine & dose no.)	**อายุที่แนะนำให้ (Recommended age)**	**อายุน้อยที่สุดของโด๊สนี้ (minimum age)**	**ระยะห่างที่แนะนำกับโด๊สถัดไป (Recommended interval)**	**ระยะห่างที่น้อยที่สุดของโด๊สถัดไป (minimum interval)**
BCG	แรกเกิด	แรกเกิด	-	-
HBV-1	แรกเกิด	แรกเกิด	1-4 เดือน	4 สัปดาห์
HBV-2[2]	1-2 เดือน	4 สัปดาห์	2-17 เดือน	8 สัปดาห์
HBV-3[3]	6-18 เดือน	24 สัปดาห์	-	-
DTwP,DTaP-1	2 เดือน	6 สัปดาห์	2 เดือน	4 สัปดาห์
DTwP,DTaP-2	4 เดือน	10 สัปดาห์	2 เดือน	4 สัปดาห์
DTwP,DTaP-3[4]	6 เดือน	14 สัปดาห์	12 เดือน	6 เดือน
DTwP,DTaP-4	18 เดือน	12 เดือน	3 ปี	6 เดือน
DTwP,DTaP-5	4-6 ปี	4 ปี	-	-

วัคซีนและโด๊สที่ให้ (vaccine & dose no.)	อายุที่แนะนำให้ (Recommended age)	อายุน้อยที่สุดของโด๊สนี้ (minimum age)	ระยะห่างที่แนะนำกับโด๊สถัดไป (Recommended interval)	ระยะห่างที่น้อยที่สุดของโด๊สถัดไป (minimum interval)
Tdap[5]	4-6 ปี ,≥11 ปี	4 ปี, 7 ปี	-	-
Td	11-12 ปี	7 ปี	10 ปี	5 ปี
OPV,IPV-1	2 เดือน	6 สัปดาห์	2 เดือน	4 สัปดาห์
OPV,IPV-2	4 เดือน	10 สัปดาห์	2 เดือน	4 สัปดาห์
OPV,IPV-3	6 เดือน	14 สัปดาห์	12 เดือน	6 เดือน
OPV,IPV-4[6]	18 เดือน	12 เดือน	3 ปี	6 เดือน
OPV,IPV-5	4-6 ปี	4 ปี	-	-
MMR-1	9-12 เดือน	9 เดือน	1½ - 5 ปี	4 สัปดาห์
MMR-2[7]	2½ - 6 ปี	1½ ปี	-	-
Inactivated JE-1	12-18 เดือน	9 เดือน	4 สัปดาห์	1 สัปดาห์
Inactivated JE-2	13-19 เดือน	10 เดือน	11 เดือน	3 สัปดาห์
Inactivated JE-3	24-30 เดือน	21 เดือน	-	-
Live JE-1	9-12 เดือน	9 เดือน	3-12 เดือน[8]	3-12 เดือน[8]
Live JE-2	12-24 เดือน	12-21 เดือน[8]	-	-
Hib-1[9]	2 เดือน	6 สัปดาห์	2 เดือน	4 สัปดาห์
Hib-2	4 เดือน	10 สัปดาห์	2 เดือน	4 สัปดาห์
Hib-3	6 เดือน	14 สัปดาห์	6-12 เดือน	8 สัปดาห์
Hib-4[10]	18 เดือน	12 เดือน	-	-
PCV-1[9]	2 เดือน	6 สัปดาห์	2 เดือน	4 สัปดาห์
PCV-2	4 เดือน	10 สัปดาห์	2 เดือน	4 สัปดาห์
PCV-3[11]	6 เดือน	14 สัปดาห์	6 เดือน	8 สัปดาห์
PCV-4	12-15 เดือน	12 เดือน	-	-
PS23-1	-	2 ปี	5 ปี	5 ปี
PS23-2	-	7 ปี	-	-
Rota-1[12]	2 เดือน	6 สัปดาห์	2 เดือน	4 สัปดาห์
Rota-2	4 เดือน	10 สัปดาห์	2 เดือน	4 สัปดาห์
Rota-3(เฉพาะ Rptateq®)	6 เดือน	14 สัปดาห์	-	-
Var-1	12-18 เดือน	12 เดือน	3-5 ปี	12 สัปดาห์
Var-2[13]	4-6 ปี	15 เดือน	-	-

วัคซีนและโด๊สที่ให้ (vaccine & dose no.)	อายุที่แนะนำให้ (Recommended age)	อายุน้อยที่สุดของโด๊สนี้ (minimum age)	ระยะห่างที่แนะนำกับโด๊สถัดไป (Recommended interval)	ระยะห่างที่น้อยที่สุดของโด๊สถัดไป (minimum interval)
HAV-1 HAV-2[15]	12-23 เดือน ≥ 18 เดือน	12 เดือน 18 เดือน	6-12 เดือน -	6 เดือน -
Influenza	≥ 6 เดือน	6 เดือน	1 ปี[14]	4 สัปดาห์
HPV-1 HPV-2[15] HPV-3	11-12 ปี 11-12 ปี (+1-2 เดือน) 11-12 (+6 เดือน)	9 ปี 9 ปี + 4-8 สัปดาห์ 9 ปี + 24 สัปดาห์	1-2 เดือน 4-5 เดือน -	4 สัปดาห์ 12 สัปดาห์ -

ตาราง 2-4 ตารางการฉีดวัคซีน ตั้งแต่อายุ 0 – 6 ปี

[1] กรณีวัคซีนรวม อายุน้อยที่สุดในการให้วัคซีน ให้ยึดอายุน้อยที่สุดของวัคซีนที่เป็นส่วนประกอบที่มากที่สุด และสำหรับระยะห่างที่น้อยที่สุดของโดสถัดไปให้ยึดระยะห่างน้อยที่สุดของวัคซีนที่เป็นส่วนประกอบมากที่สุด

[2] ถ้ามารดามี HBsAg บวก และทารกไม่ได้ HBIG ควรได้ HBV โดสที่สองที่อายุ 1 เดือน

[3] กรณีให้เป็นวัคซีนรวม DTP+HBV อาจได้ HBV ที่ 4 เดือนด้วย แต่โดสสุดท้ายของ HBV ไม่ควรก่อนอายุ 24 สัปดาห์

[4] ระยะห่างที่น้อยที่สุดของ DTwP และ DTaP โดส 3 และ 4 อย่างน้อยต้อง 6 เดือน อย่างไรก็ตามไม่จำเป็นต้องให้ DTwP และ DTaP โดส 4 ซ้ำ ถ้าได้ห่างจาก DTwP, DTaP-3 อย่างน้อย 4 เดือน

[5] Tdap ใช้แทน Td ในเด็กโตและผู้ใหญ่ได้ 1 โดส และอาจใช้แทน DTwPที่อายุ 4-6 ปี ได้

[6] ถ้าใช้ IPV อย่างเดียวตลอด อาจให้เพียง 4 โดส โดยงดโดสที่ 4 ได้

[7] อาจใช้วัคซีนรวม หัด คางทูม หัดเยอรมัน และอีสุกอีใสแทน

[8] วัคซีนเชื้อมีชีวิต (live JE) มี 2 ชนิด คือ CD-JEVAX® เริ่มฉีดที่อายุ 9-12 เดือน และ โดสที่ 2 อีก 3-12 เดือนต่อมา อีกชนิด คือ IMOJEV® เริ่มฉีดที่อายุ 9-12 เดือน และโดสที่ 2 อีก 12-24 เดือนต่อมา

[9] Hib และ PCV ถ้าเริ่มให้ที่อายุ มากกว่า 7 เดือน จำนวนโดสจะลดลง

[10] Hib-4 อาจไม่จำเป็นต้องฉีดกระตุ้นในเด็กไทยปกติ

[11] PCV-3 ในเด็กปกติอาจไม่จำเป็นต้องให้ (เป็นการฉีดแบบ 2+1)

[12] Rota โดสแรกให้อายุไม่เกิน 15 สัปดาห์ โดสสุดท้ายอายุไม่เกิน 8 เดือน

[13] Varicella vaccine ในเด็กอายุ 1-12 ปี ให้ 1-2 โดส โดสที่ 2 อาจพิจารณาฉีดที่อายุ 4-6 ปี ในกรณีมีการระบาดอาจฉีดโดสที่สองก่อนอายุ 4 ปี แต่ต้องหางจากโดสแรกอย่างน้อย 3 เดือน ถ้าให้ในเด็กอายุ >13 ปี ให้ 2 โดส โดสที่สองห่างจากโดสแรก 4 สัปดาห์

[14] Influenza vaccine ใน เด็กอายุน้อยกว่า 9 ปี ถ้าไม่เคยได้วัคซีนมาก่อน ในปีแรกให้ 2 โดส ห่างกัน 1 เดือน

[15] HPV ชนิด 4 สายพันธุ์ (Gardasil®) ให้ฉีด 3 โดส (0, 2, 6 เดือน) แนะนำในหญิงและชาย 9-26 ปี แต่ในวัยรุ่นที่แข็งแรงดีหากฉีดเข็มแรกก่อนอายุ 15 ปี ให้ฉีด 2 เข็มได้ที่ 0, 6-12 เดือน

Hib หากเริ่มฉีดยาช้า ให้พิจารณาตามตาราง

อายุที่เริ่มฉีด	**เดือนที่ของการฉีด Hib**
2-6 เดือน	0, 2, 4, ฉีดกระตุ้นที่ 12-18 เดือน
7-11 เดือน	0, 2, ฉีดกระตุ้นที่ 12-18 เดือน
12-24 เดือน	เข็มเดียว
> 24 เดือน เฉพาะผู้ที่เสี่ยง*	0, 2

ตาราง 2-5 ตารางการฉีดวัคซีนฮิบ

*ผู้ที่เสี่ยงต่อโรคฮิบ เช่น ผู้ที่มีภูมิคุ้มกันบกพร่อง ไม่มีม้าม หรือม้ามทำงานผิดปกติ

PCV (Prevnar13®) หากเริ่มฉีดยาช้า ให้พิจารณาตามตาราง

อายุที่เริ่มฉีด	**จำนวนครั้งที่ฉีด**	**การฉีดกระตุ้น**
เด็กปกติและเด็กเสี่ยง 2-6 เดือน	3 ครั้ง ห่างกัน 6-8 สัปดาห์	1 ครั้ง อายุ 12-15 เดือน
เด็กปกติและเด็กเสี่ยง 7-11 เดือน	2 ครั้ง ห่างกัน 6-8 สัปดาห์	1 ครั้ง อายุ 12-15 เดือน

อายุที่เริ่มฉีด	จำนวนครั้งที่ฉีด	การฉีดกระตุ้น
เด็กปกติและเด็กเสี่ยง 12-23 เดือน	2 ครั้ง ห่างกัน 6-8 สัปดาห์	ไม่ต้องฉีด
เด็กปกติ 2-5 ปี	เข็มเดียว	ไม่ต้องฉีด
เด็กเสี่ยง* - อายุ 2-6 ปี - อายุ >6-18 ปี	 2 ครั้ง ห่างกัน 6-8 สัปดาห์ 1 ครั้ง	 ฉีดกระตุ้นด้วย PS-23 1 เข็ม ห่างจาก PCV เข็มสุดท้าย 8 สัปดาห์$

ตาราง 2-6 ตารางฉีดวัคซีน PCV

*เด็กเสี่ยง คือ เด็กที่มีโอกาสเป็นโรคติดเชื้อ PCV อย่างรุนแรงมากกว่าเด็กปกติ ได้แก่ เด็กที่เป็นโรคภูมิคุ้มกันบกพร่อง ภาวะไม่มีม้าม ธาลัสซีเมีย โรคเรื้อรังของอวัยวะต่างๆ เช่นโรคปอด รวมทั้งหอบหืดรุนแรง โรคหัวใจ โรคตับ โรคไต โรคเบาหวาน และโรคที่เสี่ยงต่อเยื่อหุ้มสมองอักเสบ เช่น CSF leak, cochlear implantation

$เด็กเสี่ยงประเภท ภูมิคุ้มกันบกพร่อง ภาวะไม่มีม้าม หรือธาลัสซีเมีย ควรฉีด PS-23 ซ้ำอีก 1 ครั้ง ห่างจาก PS-23 ครั้งแรก 5 ปี

2.2 **การเรียนรู้ของเครื่องจักร (Machine Learning)**

Machine Learning คือ การเรียนรู้ของเครื่องจักร เป็นระบบที่สามารถเรียนรู้ได้จากตัวอย่างด้วยตนเอง โดยปราศจากการป้อนคำสั่งของโปรแกรมเมอร์ ความก้าวหน้าในครั้งนี้ มาพร้อมกับความคิดที่ว่า เครื่องคอมพิวเตอร์สามารถเรียนรู้เพียงแค่จากข้อมูลอย่างเดียว เพื่อที่จะผลิตผลลัพธ์ที่แม่นยำออกมาได้ รูปแบบการเรียนรู้ของเครื่องจักรสามารถแบ่งได้ 3 แบบ ดังนี้

รูปแบบที่ 1 การเรียนรู้แบบมีผู้สอนหรือมีผู้แนะนำ (Supervised Learning) เป็นการแยกประเภท หรือบอกผลลัพธ์ที่ควรจะเป็นไว้ล่วงหน้า จากนั้นจึงใช้ข้อมูลตัวอย่างฝึกสอน แล้วนำไปผ่านอัลกอริทึมสำหรับสร้างโมเดล ที่ไว้สำหรับคาดการณ์ผลลัพธ์ที่จะเกิดขึ้นในอนาคต เมื่อเกิดการเรียนรู้แล้ว สามารถนำข้อมูลใหม่ที่เครื่องมือไม่เคยเห็น ทำการใส่ข้อมูลเข้าไปเครื่องมือ Machine Learning จะสามารถคาดการณ์ผลลัพธ์ที่เกิดขึ้นได้ ใน Supervised Learning มี 2 ประเภท คือ การแบ่งแยกประเภท (Classification) และการถดถอย (Regression)

- **การแบ่งแยกประเภท (Classification)**

การแบ่งแยกประเภท เป็นการจำแนกประเภทข้อมูล ของ Machine Learning แบบ Supervised Learning โดยมีตัวอย่างในชุดข้อมูลสอน (training set) ที่ใช้ จะมีคุณลักษณะหนึ่งซึ่งบอกคำประเภทของตัวอย่างนั้น เรียกว่าค่าคุณลักษณะนี้ว่าฉลากบอกประเภท (class label) ซึ่งเป็นค่าข้อมูลแบบ categorical ยกตัวอย่างเช่น ต้องการทำนายเพศของลูกค้าคนหนึ่ง ต้องเก็บข้อมูลตั้งแต่ชื่อ ส่วนสูง น้ำหนัก อาชีพ เป็นต้น จากฐานข้อมูลลูกค้า สามารถตอบได้เพียงว่าลูกค้าผู้ชายหรือผู้หญิงเท่านั้น ผลลัพธ์จาก ตัวแบ่งแยกประเภท(classifier) นี้จะถูกกำหนดด้วยความน่าจะเป็นว่าเป็นผู้ชายหรือผู้หญิง (คือ การทำสัญลักษณ์ไว้ (label)) โดยขึ้นอยู่กับข้อมูลเป็นหลัก(ลักษณะ (feature) ที่เราเก็บมา) เมื่อ model เรียนรู้ที่จะจดจำว่าเป็นผู้ชายหรือผู้หญิง จะสามารถนำข้อมูลใหม่มาทำนายได้ ยกตัวอย่างเช่น ได้รับข้อมูลใหม่จากลูกค้าใหม่และต้องการทราบว่าลูกค้าคนนั้นเป็นผู้ชายหรือผู้หญิง ถ้า classifier นี้ทำนายว่ามีโอกาสเป็นผู้ชาย 70% นั่นหมายความว่าอัลกอริทึมมีความแน่นอนอยู่ที่ 70% ที่ลูกค้าคนนั้นจะเป็นผู้ชายและ 30% ที่จะเป็นผู้หญิง

- **การถดถอย (Regression)**

เมื่อผลลัพธ์มีค่าต่อเนื่องกัน จะเป็นงานของ regression ที่จะมาช่วยแก้ปัญหานี้ ยกตัวอย่างเช่น นักวิเคราะห์ด้านการเงินคนหนึ่งอาจต้องการทำนายมูลค่าของหุ้นโดยดูจาก feature ต่าง ๆ เช่น ส่วนได้ส่วนเสีย (equity) สถานะของหุ้นก่อนหน้านี้ และดัชนีเศรษฐกิจมหภาค ระบบจะถูก train เพื่อประเมินราคาของหุ้นด้วยความผิดพลาดที่น้อยที่สุด

รูปแบบที่ 2 การเรียนรู้แบบไม่มีผู้สอนหรือไม่มีผู้แนะนำ (Unsupervised Learning) เป็นการเรียนรู้ของเครื่องมือที่ไม่สามารถทราบคำตอบได้ชัดเจน ซึ่งเราต้องการให้เครื่องมือหาวิธีสร้างโครงสร้างข้อมูลที่เราไม่รู้จัก (Unknown Structure)

รูปแบบที่ 3 การเรียนรู้แบบเสริมแรง (Reinforcement Learning) เป็นเครื่องมือ Artificial Intelligence (AI) มากที่สุด เป็นการเรียนรู้และเปลี่ยนไปตามสิ่งแวดล้อม ตัวอย่าง กรณีเป็นข่าวไปทั่วโลก คือ ระบบ AI ชื่อว่า “Alpha Go” ที่พัฒนาโดยบริษัท Deep Mind ของ Google นั้น โดย Alpha Go ได้ทำการพัฒนาและศึกษารูปแบบการเรียนของเซียนโกะทั่วโลก เพื่อพัฒนาเป็นรูปแบบการเล่นของตัวเอง มีการฝึกฝนเพื่อเพิ่มความสามารถ โดยฝึกกับ

ตัวเองในรูปแบบการเล่นต่างๆนับล้านครั้ง ทำให้ฉลาดและเล่นเก่งขึ้นทุกครั้ง และ สามารถชนะเซียนโกะระดับโลกที่มีฝีมือขั้นสูงสุดในการเล่นโกะได้

2.3 **K-Nearest Neighbors**

K-Nearest Neighbors (K-NN) คือ เป็นขั้นตอนวิธีในการคำนวณหาระยะห่างระหว่างจุดของข้อมูลที่สนใจกับทุกจุดข้อมูลทั้งหมด เมื่อได้ระยพห่างระหว่าจุดทุกจุดกับข้อมูลจุดที่สนใจแล้ว ทำการเลือกข้อมูลจุดที่ใกล้ที่สุดกับจุดข้อมูลที่สนใจ K ตัว โดยขั้นตอนวิธี K- Nearest Neighbors มีขั้นตอน ดังนี้

1. กำหนดขนาดของ K
2. คำนวณระยะห่างของข้อมูลที่ต้องพิจารณากับกลุ่มข้อมูล
3. จัดเรียงลำดับของระยะห่าง และเลือกพิจารณาชุดข้อมูลที่ใกล้กับจุดที่สนใจพิจารณาจากตามจำนวน K ที่กำหนดไว้
4. พิจารณาข้อมูลจำนวน K ชุด และสังเกตว่าจุดไหนใกล้กับจุดที่สนใจมากที่สุด

2.4 **ระบบแนะนำ (Recommender System)**

ระบบแนะนำ คือ ระบบสนับสนุนการตัดสินใจที่ให้การแนะนำสินค้า หรือผลิตภัณฑ์ หรือบริการที่มีความเหมาะสม กับรูปแบบและพฤติกรรมของลูกค้าแต่ละคน โดยอาศัยข้อมูลของผู้ใช้งานร่วมกับข้อมูลประกอบภายนอก มาใช้ในการวิเคราะห์คัดกรอง ให้ได้สิ่งที่มีความเหมาะสมต่อผู้ใช้งาน ระบบแนะนำ มีรูปแบบดังต่อไปนี้

2.3.1 **Collaborative Filtering**

เป็นการพิจารณาความชอบที่มีต่อไอเทมเป็นความชอบของผู้ใช้ เช่น การให้คะแนน (Rating) จะนำไปพิจารณากับไอเทมที่ผู้ใช้ที่ยังไม่ให้ความนิยมไว้ แต่มีความคล้ายกับไอเทมที่ผู้ใช้เคยให้คะแนนความชอบไว้แล้วในอดีต แบ่งออกเป็น 2 ประเภท ดังนี้

2.3.1.1 User-Based Filtering เป็นการกรองข้อมูลร่วม โดยพิจารณาจากผู้ใช้ที่พฤติกรรมเหมือนกัน

2.3.1.2 Item-Based Filtering เป็นการกรองข้อมูลร่วม โดยพิจารณาจากรายการสินค้าที่ถูกซื้อด้วยลูกค้ากลุ่มเดียวกัน

วิธีในการแนะนำต้องพิจารณาหาความคล้ายของผู้ใช้เป้าหมายกับผู้ใช้คนอื่นที่อยู่ในฐานข้อมูล โดยใช้สมการที่ (2.1)

วิธีการหาค่าความคล้ายคลึงแบบโคไซน์ (Cosine-based Similarity)

$$sim(u_i, u_v) = \frac{\sum j(r_{u_i,m_j} r_{u_v,m_j})}{\sqrt{\sum j(r_{u_i,m_j})^2}\sqrt{\sum j(r_{u_v,m_j})^2}} \quad (2.1)$$

โดยที่

U = $\{u_1, u_2, \dots, u_i, \dots, u_{|U|}\}$ คือ เซตของผู้ใช้

M = $\{m_1, m_2, \dots, m_i, \dots, m_{|U|}\}$ คือ เซตของไอเทม

r_{u_v,m_j} คือ คะแนนไอเทม m_j ที่ถูกให้คะแนนโดยผู้ใช้ u_v

r_{u_i,m_j} คือ คะแนนไอเทม m_j ที่ถูกให้คะแนนโดยผู้ใช้ u_i

ข้อดีของ Collaborative Filtering คือ สามารถประยุกต์ใช้ได้กับทุกประเภทของสินค้าและไม่ต้องมีการทำแคตาล็อคเพื่ออธิบายสินค้า แต่จะมีข้อจำกัดเมื่อข้อมูล rating มีน้อยหรือสินค้าใหม่ๆ ที่ยังไม่ค่อยมีการให้ rating จะแนะนำได้ยาก

2.3.2 **Content-based Filtering**

Content-based Filtering จะดูที่ลักษณะของสินค้าที่จะแนะนำ และแนะนำสิ่งที่มีลักษณะหรือมีคำอธิบายคล้ายกับโปรไฟล์ของผู้ใช้ รวมถึงลักษณะของสิ่งที่ผู้ใช้เคยใช้หรือเคยชอบ เช่น ระบบจะแนะนำหนังที่เนื้อหาของหนังมีความคล้ายกับหนังที่ผู้ใช้เคยดูมาก่อนหน้านี้ ดังนั้นการใช้ Content-based Filtering จะต้องมีข้อมูลคุณลักษณะของสินค้า เช่น กลุ่ม/ประเภท ขนาด ราคา สี สไตล์ ฟังก์ชั่น ตำแหน่งแบรนด์ดิ้ง หรือคำอธิบายตัวสินค้าแบบย่อ เป็นต้น

ข้อดีของ Content-based Filtering มีดังนี้

- เนื่องจากระบบแนะนำดูโปรไฟล์ของผู้ใช้แต่ละคนแยกออกจากกัน สินค้าที่แนะนำจะค่อนข้างตรงกับรสนิยมของผู้ใช้ที่มีรสนิยมแตกต่างจากคนส่วนใหญ่
- การแนะนำสินค้าใหม่ที่ยังไม่ค่อยมีผู้ใช้งานจะทำได้ง่ายเพราะสามารถพิจารณาจากความคล้ายคลึงของคุณลักษณะกับสินค้าเดิม

ความยากของการทำระบบแนะนำแบบ Content-based Filtering คือ การเตรียมแคตาล็อคข้อมูลสินค้าซึ่งใช้เวลามาก และการสร้าง feature ที่เหมาะสมเพื่ออธิบายตัวสินค้าและโปรไฟล์ของผู้ใช้ซึ่งขึ้นอยู่กับแนวของสินค้า นอกจากนั้น Content-based Filtering จะไม่สามารถแนะนำสินค้าที่แตกต่างจากสินค้าที่ผู้ใช้เคยซื้อมากนัก ทำให้ผู้ใช้ได้รับการแนะสินค้าที่มีความหลากหลายค่อนข้างน้อย

2.5 **Gradient Boosting Classifier**

Gradient Boosting Classifier เป็น Boosting ขั้นตอนวิธี ใน Ensemble Learning ที่ใช้ Classifier หลายข้อมูลมาช่วยสร้างโมเดลและพยากรณ์ทำงานเป็นโซ่ต่อกัน โดยแต่ละตัวจะแก้ไขจุดด้อยของ Classifier ตัวก่อนหน้า เมื่อ Training เสร็จแล้ว Classifier ทุกตัวจะพยากรณ์ร่วมกัน Ensemble Learning จะเป็นการรวมของโมเดลการเรียนรู้ที่หลากหลายที่มีความแตกต่างและมีอิสระต่อกัน เพื่อเพิ่มประสิทธิภาพของโมเดล

2.6 **งานวิจัยที่เกี่ยวข้อง**

Farag Hamad Kuwil และคณะ (2020) ได้นำเสนองานวิจัยเรื่อง A novel data clustering algorithm based on gravity center methodology งานวิจัยได้นำเสนอการแนะนำอัลกอริธึมการจัดกลุ่มข้อมูลแบบใหม่โดยใช้วิธีการศูนย์แรงโน้มถ่วง จุดแข็งของขั้นตอนวิธีศูนย์แรงโน้มถ่วง คือ ไม่ต้องระบุพารามิเตอร์ แต่ข้อเสีย คือ ไม่สามารถทำงานได้ดีกับข้อมูลที่มีปริมาณน้อยมาก ผลการจัดกลุ่มข้อมูลโดยใช้ Gravity Center เปรียบเทียบกับวิธีอื่นๆ อีก 3 วิธี ได้แก่ K-mean, K-medians และ K-medoids ผลการวิจัยระบุว่า Gravity Center มีประสิทธิภาพเหนือกว่าขั้นตอนวิธีเหล่านั้นด้วยชุดข้อมูลที่ใช้ ได้แก่ NNDSS, Health-infectiousdisease-2001 2014, Unplanned Hospital Visits-Hospital, Diabetes และ Medicare National DMEPOS HCPCS

Jiawei Yang และคณะ (2021) ได้นำเสนองานวิจัยเรื่อง Mean-shift outlier detection and filtering งานวิจัยได้นำเสนอเทคนิคการตรวจหาและกรองค่าเบี่ยงเบนเฉลี่ย เพื่อกำจัดความเอนเอียงที่เกิดจากค่าผิดปกติ ขั้นตอนวิธีนี้ใช้ได้กับทั้งข้อมูลตัวเลขและสตริง ผลการทดลองของวิธีนี้สำหรับงานนี้มีประสิทธิภาพดีกว่าขั้นตอนวิธีการลบค่าผิดปกติอื่นๆ อีก 5 วิธี ได้แก่ LOF, ODIN, NC, IFOREST และ ABOD ขั้นตอนวิธีนี้ยังมีประสิทธิภาพเหนือกว่าวิธีการตรวจหาค่าผิดปกติที่มีอยู่ ได้แก่ LOF, NC, KNN, ODIN, MCD, IFOREST, OCSVM, PCAD และ ABOD ในงานนี้ใช้ชุดข้อมูลในการทดลอง คือ KDD-Cup99, Stamps, PageBlocks, Pima, Arrhythmia และ Parkinson

Xiaokang Wang และคณะ (2020) เสนองานวิจัยเรื่อง A density weighted fuzzy outlier clustering approach for class imbalanced learning ซึ่งเป็นวิธีการจัดกลุ่มค่าผิดปกติแบบคลุมเครือ วิธีนี้จะพิจารณาความสัมพันธ์ของพื้นที่ใกล้เคียงที่คลุมเครือใหม่กับข้อมูลความหนาแน่น เมื่อน้ำหนักข้อมูลตัวอย่างในกระบวนการจัดกลุ่มถูกผสมกับวิธีการจัดกลุ่มค่าผิดปกติแบบคลุมเครือ ด้วยวิธีนี้ ตัวอย่างที่เป็นตัวแทนส่วนใหญ่จะถูกเลือก แต่ตัวอย่างที่ผิดปกติจะถูกกำจัดออกไป ความถูกต้องแม่นยำของวิธีนี้แสดงให้เห็นประสิทธิภาพที่เหนือกว่าถึง 92% เมื่อเทียบกับโมเดลการสุ่มตัวอย่างคลัสเตอร์อื่นๆ วิธีการ density-weighted fuzzy outlierจัดกลุ่มค่าผิดปกติที่ถ่วงน้ำหนักโดยความหนาแน่นสามารถใช้กับปัญหาที่ไม่สมดุลในชีวิตจริงได้ งานวิจัยนี้ใช้ชุดข้อมูล Blood-transfusion, Parkinson, Sick_numeric2, WDBC และ Wine

Jiang Xie และคณะ(2020) ได้นำเสนองานวิจัยเรื่อง A local-gravitation-based method for the detection of outliers and boundary points งานวิจัยได้นำเสนอ จุดข้อมูลแต่ละจุดจะถูกมองว่าเป็นวัตถุที่มีทั้งมวลและแรงผลลัพธ์ในพื้นที่ (LRF) ที่สร้างโดยเพื่อนบ้าน เมื่อจำนวนเพื่อนบ้านเพิ่มขึ้น LRF ของค่าผิดปกติ จุดขอบเขต และคะแนนภายในจะเปลี่ยนแปลงในอัตราที่แตกต่างกัน อัตราการเปลี่ยนแปลง LRF ของจุดความหนาแน่นต่ำกว่ามีคะแนนสูงกว่า นั่นคือ อัตราการเปลี่ยนแปลงของค่าผิดปกติจะสูงกว่าอัตราของขอบเขตและจุดภายใน กล่าวอีกนัยหนึ่ง คะแนนอันดับสูงสุดสามารถระบุได้ว่าเป็นค่าผิดปกติ ในทำนองเดียวกัน ยิ่งอัตราการเปลี่ยนแปลงของจุดสูงขึ้นเท่าใด LRF ก็จะยิ่งสูงขึ้น และโอกาสที่จุดนั้นจะเป็นจุดขอบเขตก็จะยิ่งมากขึ้น ข้อได้เปรียบหลักของวิธีนี้คือไม่ขึ้นกับการเลือกค่า K-value ซึ่งจะส่งผลให้ประสิทธิภาพในการตรวจจับดีขึ้น ใช้ชุดข้อมูล Heart disease, Lymphography, Ionosphere, Breast cancer Wisconsin, Blood transfusion service center และ SPECTF

Aditya Hari Bawono และคณะ (2019) ได้นำเสนองานวิจัยเรื่อง Outlier Detection with Supervised Learning Method งานวิจัยได้นำเสนอ วิธีการเป็นที่นิยมหลายวิธี ได้แก่ K-Nearest Neighbor, Centroid Classi er และ Naive Bayes ถูกนำมาเปรียบเทียบเป็นเครื่องมือในการจัดการงานการตรวจจับค่าผิดปกติ ผลการวิจัยพบว่าวิธีการเหล่านี้ได้ผลลัพธ์ 81% สำหรับการวิจัยในอนาคตเพื่อปรับวิธีการดังกล่าวเพื่อปรับปรุงประสิทธิภาพ Elhossinyและคณะ เสนอโดยใช้ K-Means++ ที่ปรับปรุงเพื่อจัดการกับข้อมูลที่ขาดหายไปและค่าผิดปกติ ผลการทดลองโดยใช้ตัววัด RMSE ได้ 17% ใช้ชุดข้อมูล Thyroid, Vertebral, Wine, Satellite, Breast cancer และ Ionosphere

Diego Luchi และคณะ (2019) ได้นำเสนองานวิจัยเรื่อง Sampling approaches for applying DBSCAN to large datasets งานวิจัยนำเสนอ DBSCAN เป็นวิธีการจัดกลุ่มสำหรับการระบุคลัสเตอร์ที่ต่างกันและแยกเดี่ยวที่มีสัญญาณรบกวน มีบทความมากมาย ในการจัดการกับปัญหา

การปรับขนาด DBSCAN แม้จะมีปัญหาเรื่องความสามารถในการปรับขนาด แต่อัลกอริธึม DBSCAN สามารถลดเวลาในการดำเนินการได้เนื่องจากจำนวนรูปแบบข้อมูลลดลง และยังเสนอฮิวริสติกใหม่ที่เรียกว่า I-DBSCAN ซึ่งสามารถแก้ไขและสร้างผลลัพธ์ที่ดีสำหรับชุดข้อมูลทั้งหมดโดยไม่มีพารามิเตอร์เพิ่มเติม ในงานนี้ใช้ชุดข้อมูล Abalone (Scale), Mushrooms, Pendigits, Letter, Cadata, Shuttle, Sensorless (Scale), SensIT (acoustic), SensIT (seismic), Skin-Nonskin และ Poker

Pawel Karczmarek และคณะ (2021) ได้นำเสนองานวิจัยเรื่อง Fuzzy C-Means-based Isolation Forest งานวิจัยได้นำเสนอ การวิเคราะห์ในรายละเอียดสำหรับลักษณะเฉพาะของข้อมูลตัวอย่างเช่น พิจารณาขนาดฐานข้อมูล จำนวนแอตทริบิวต์ของระเบียน และชนิดข้อมูล FCM ใช้ขั้นตอนวิธี Isolation Forest ที่สร้างขึ้นไปยังคลัสเตอร์ตามโหนดเหล่านี้จะถูกสร้างขึ้น ดังนั้น สิ่งนี้สามารถนำไปสู่ข้อบกพร่อง สำหรับองค์ประกอบที่กำหนดซึ่งอาจอยู่ในกลุ่มขององค์ประกอบที่คล้ายคลึงกัน เพื่อแก้ปัญหานี้ จึงมีการนำเสนอวิธี Fuzzy C-Means ผลการทดลองดำเนินการกับชุดข้อมูล 27 ชุด พบว่า FCM สามารถมีบทบาทสำคัญในการปรับปรุงประสิทธิภาพการทำงานของวิธีการตรวจหาความผิดปกติ ใช้ข้อมูล ได้แก่ Annthyroid, Arrhythmia, Breastw, Cardio, Cover, Glass, Ionosphere, Letter, Lympho, Mammography, Mnist, Musk, Optdigits, Pendigits, Pima, ดาวเทียม, Satimage-2, Shuttle, Speech, Thyroid, Vertebral, Vowels, Wine และ Nad

Pawel Karczmarek และคณะ (2020) ได้นำเสนองานวิจัยเรื่อง K-Means-based isolation forest งานวิจัยได้นำเสนอ ขั้นตอนวิธี Isolation Forest แบบ K-Means ช่วยให้สามารถสร้างแผนที่การค้นหาได้ วิธีการที่เสนอนี้เป็นประโยชน์สำหรับข้อมูลที่มาจากพื้นที่การใช้งานต่างๆ รวมทั้งการขนส่งระหว่างแบบจำลองและข้อมูลเชิงพื้นที่ ข้อดีของวิธีนี้ คือ สามารถป้อนข้อมูลในกระบวนการสร้างแผนผังการตัดสินใจได้ นอกจากนี้ยังส่งกลับคะแนนความผิดปกติที่น่าสนใจยิ่งขึ้นอีกด้วย ใช้ชุดข้อมูล NYC Taxi, NYC Taxi (geographical positions), Ship transport และ Train transport

Raid Lafta และคณะ (2015) ได้นำเสนองานวิจัยเรื่อง An Intelligent Recommender System based on Short-term Risk Prediction for Heart Disease Patients งานวิจัยได้นำเสนอพัฒนาระบบแนะนำโรคหัวใจ ใช้ทำนายช่วงเวลา เพื่อให้คำแนะนำแก่ผู้ป่วยโรคหัวใจห่างไกล วิเคราะห์การทดสอบทางการแพทย์ของผู้ป่วยในแต่ละรายที่บันทึก โดยใช้ข้อมูลทางการแพทย์ เช่น อัตราการเต้นของหัวใจ ผลลัพธ์เท่ากับ 0 คือ มีความเสี่ยงต่ำ ส่วน 1คือ มีความเสี่ยงสูง แล้วแนะนำ ถ้าเป็น 0 ไม่ต้องทดสอบ ถ้าเป็น 1 ต้องทำการทดสอบเพิ่มเติม เพื่อสนับสนุนการตัดสินใจแก่ผู้ป่วย

Reena Pagare และคณะ (2012) นำเสนองานวิจัยเรื่อง A Study of Recommender System Techniques งานวิจัยได้นำเสนอกล่าวถึงการค้นหารายละเอียดของสินค้าบนเว็บไซต์จาก

ผู้ใช้หรือผู้เชี่ยวชาญที่ทำการ Review ของสินค้า การ Review ที่ผู้ใช้ให้ความคิดเห็น ดังนั้นระบบแนะนำจึงมีความสำคัญ ทำให้สามารถตอบสนองกับปัญหาข้อมูลที่มีปริมาณมาก โดยงานวิจัยนี้ได้นำเสนอขั้นตอนวิธี Collaborative Filtering ในการสร้างระบบผู้แนะนำที่มีคุณภาพสูงของการแนะนำของผู้ใช้ที่มีความชอบคล้ายกัน นอกจากนี้ ในงานวิจัยนี้ยังได้มีการเปรียบเทียบ ขั้นตอนวิธีดังนี้ : Content Based Method, Collaborative filtering Method, The user-based algorithm, The item-based algorithm, The dimensionality-reduction algorithm, The generative-model algorithm, The spreading-activation algorithm, The link-analysis algorithm, Trust-Based Method สามารถสรุปได้ว่า Content-Based ต้องมีข้อมูลขอบเขตที่ชัดเจน ส่วน Collaborative Filtering จำเป็นต้องมี Rating สินค้าจากผู้ใช้ ขั้นตอนวิธี Trust Based จะสามารถใช้แก้ปัญหา Cold start หรือปัญหาข้อมูลมีจำนวนน้อยได้ ในขั้นตอนวิธี 6 วิธี ที่กล่าวถึงข้างต้น Link-analysis Algorithm ทำงานได้ดีกว่าขั้นตอนวิธีอื่นๆ

Jain Sarika และคณะ (2015) นำเสนองานวิจัยเรื่อง Trends, Problems And Solutions of Recommender System งานวิจัยได้นำเสนอ ระบบแนะนำต่างๆ เช่น Content-based, Collaborative filtering, Hybrid collaborative filtering เป็นต้น เพื่อต้องการแก้ปัญหาขั้นตอนวิธีที่พบในระบบแนะนำ โดยปัญหาที่พบในระบบแนะนำ ได้แก่

- ปัญหา Cold-start มักเกิดจากผู้ใช้ใหม่หรือมีสินค้าเพิ่มมาใหม่ โดยยังไม่มี Rating
- ปัญหา Sparsity คือ มีข้อมูลน้อยหรือมีข้อมูลที่ไม่เพียงพอในการแนะนำผู้ใช้
- ปัญหา Over-Specialization คือ ระบบแนะนำที่ดีจะต้องทำการแนะนำสินค้าที่หลากหลาย เป็นสิ่งที่ Content- Based ขาดไป

วิธีการแก้ไขปัญหา มีดังนี้

- Cold-start สามารถใช้ Collaborative filtering กับ Demographic recommending เสนอสินค้ากับผู้ใช้รายใหม่
- Sparsity สามารถใช้เทคนิค Hybrid recommendation แทนที่จะใช้ content-based แบบเดียว สามารถรวมใช้ Content-based กับ Collaborative filtering
- Over-Specialization สามารถใช้เทคนิค Collaborative filtering ข้อมูลที่ใกล้กัน

Abderrahmane Kouadria และคณะ (2019) นำเสนองานวิจัยเรื่อง A Multi-criteria Collaborative Filtering Recommender System Using Learning-to-Rank and Rank Aggregation งานวิจัยได้นำเสนอระบบแนะนำใช้วิธี Top-N สำหรับใช้พยากรณ์จัดลำดับ จากข้อมูลของคำแนะนำ ลำดับในรายการที่สร้างขึ้นมีความสำคัญมากกว่าการให้คะแนน บทความนี้เสนอระบบการจัดลำดับแบบผสม 3 ขั้นตอน สำหรับระบบแนะนำแบบหลายเกณฑ์ ขั้นตอนแรกทำการแยกเมทริกซ์รายการผู้ใช้แบบหลายเกณฑ์ออกเป็นเมทริกซ์รายการผู้ใช้ที่มีอันดับเดียว ขั้นตอนที่สองจัดลำดับ

แต่ละรายการโดยใช้ขั้นตอนวิธีเพื่อจัดลำดับ ขั้นตอนสามจะรวมรายการและจัดลำดับ ในงานวิจัยนี้ได้ทำกับข้อมูล Yahoo! Moovie

บทที่ 3

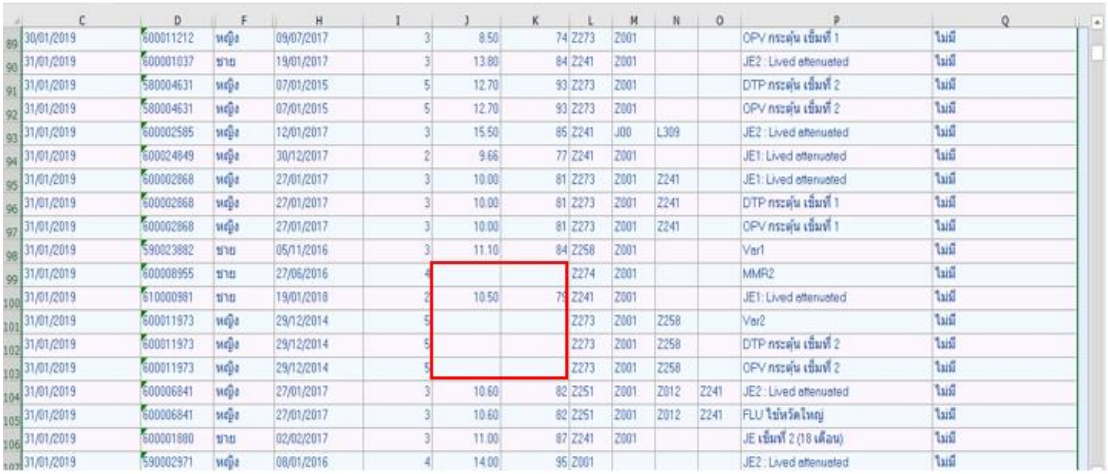
วิธีดำเนินงานวิจัย

ในงานวิจัยนี้ เป็นการศึกษาแนวทางการพัฒนาระบบแนะนำวัคซีนรายคนสำหรับเด็ก จากฐานข้อมูลประวัติการเข้ารับวัคซีน ที่จัดเก็บไว้ในฐานข้อมูลของโรงพยาบาล ซึ่งระบบนี้ จะสามารถใช้เพื่อสนับสนุนการวินิจฉัยของแพทย์ในการให้วัคซีนสำหรับเด็ก โดยงานวิจัยนี้ มีกรอบการทำงานซึ่งแบ่งการดำเนินงานออกเป็น 2 เฟส (Phase) ดังแสดงในภาพที่ 3-1

ภาพที่ 3-1 กรอบการดำเนินงานวิจัย

3.1 กรอบการดำเนินงานวิจัย

จากภาพ 3.1 กรอบการดำเนินงานวิจัย ซึ่งแบ่งออกเป็น 2 เฟส นั้น ใน Phase I จะเป็นการจัดการกับข้อมูลที่มีอยู่ในฐานข้อมูลแบบอัตโนมัติ เพื่อจัดการกับข้อมูลซึ่งมีอยู่จำนวนมาก โดยข้อมูลเหล่านี้เกิดจากการบันทึกข้อมูลทางการแพทย์ ด้วยบุคลากรทางการแพทย์หลายคน ซึ่งแต่ละคนจะมีรูปแบบในการบันทึกข้อมูลที่แตกต่างกัน หลากหลายรูปแบบ ทำให้ข้อมูลที่ควรจะอยู่ในกลุ่มเดียวกันหรือประเภทเดียวกัน ถูกประเมินว่าต่างกันซึ่งจะส่งผลให้ข้อมูลอาจมีความคลาดเคลื่อน หรืออาจนำไปใช้ประโยชน์ได้ไม่เหมาะสม หรืออาจทำให้ข้อมูลไม่ถูกต้อง ดังแสดงในภาพ 3.2



	C	D	F	H	I	J	K	L	M	N	O	P	Q
89	30/01/2019	600011212	หญิง	09/07/2017	3	8.50	74	Z273	Z001			OPV กระตุ้น เข็มที่ 1	ไม่มี
90	31/01/2019	600001037	ชาย	19/01/2017	3	13.80	84	Z241	Z001			JE2 : Lived attenuated	ไม่มี
91	31/01/2019	580004631	หญิง	07/01/2015	5	12.70	93	Z273	Z001			DTP กระตุ้น เข็มที่ 2	ไม่มี
92	31/01/2019	580004631	หญิง	07/01/2015	5	12.70	93	Z273	Z001			OPV กระตุ้น เข็มที่ 2	ไม่มี
93	31/01/2019	600002585	หญิง	12/01/2017	3	15.50	85	Z241	J00	L309		JE2 : Lived attenuated	ไม่มี
94	31/01/2019	600024849	หญิง	30/12/2017	2	9.66	77	Z241	Z001			JE1: Lived attenuated	ไม่มี
95	31/01/2019	600002868	หญิง	27/01/2017	3	10.00	81	Z273	Z001	Z241		JE1: Lived attenuated	ไม่มี
96	31/01/2019	600002868	หญิง	27/01/2017	3	10.00	81	Z273	Z001	Z241		DTP กระตุ้น เข็มที่ 1	ไม่มี
97	31/01/2019	600002868	หญิง	27/01/2017	3	10.00	81	Z273	Z001	Z241		OPV กระตุ้น เข็มที่ 1	ไม่มี
98	31/01/2019	590023882	ชาย	05/11/2016	3	11.10	84	Z258	Z001			Var1	ไม่มี
99	31/01/2019	600008955	ชาย	27/06/2016	4			Z274	Z001			MMR2	ไม่มี
100	31/01/2019	610000981	ชาย	19/01/2018	2	10.50	79	Z241	Z001			JE1: Lived attenuated	ไม่มี
101	31/01/2019	600011973	หญิง	29/12/2014	5			Z273	Z001	Z258		Var2	ไม่มี
102	31/01/2019	600011973	หญิง	29/12/2014	5			Z273	Z001	Z258		DTP กระตุ้น เข็มที่ 2	ไม่มี
103	31/01/2019	600011973	หญิง	29/12/2014	5			Z273	Z001	Z258		OPV กระตุ้น เข็มที่ 2	ไม่มี
104	31/01/2019	600006841	หญิง	27/01/2017	3	10.60	82	Z251	Z001	Z012	Z241	JE2 : Lived attenuated	ไม่มี
105	31/01/2019	600006841	หญิง	27/01/2017	3	10.60	82	Z251	Z001	Z012	Z241	FLU ไข้หวัดใหญ่	ไม่มี
106	31/01/2019	600001880	ชาย	02/02/2017	3	11.00	87	Z241	Z001			JE เข็มที่ 2 (18 เดือน)	ไม่มี
107	31/01/2019	590002971	หญิง	08/01/2016	4	14.00	95	Z001				JE2 : Lived attenuated	ไม่มี

ภาพที่ 3-2 ตัวอย่างข้อมูลที่ไม่ครบถ้วน

จากภาพที่ 3-2 ตัวอย่างข้อมูลที่ไม่ครบถ้วน ซึ่งเป็นตัวอย่างข้อมูลตั้งต้นที่ได้รับมาจากฐานข้อมูลคลินิกเด็กสุขภาพดี เพื่อใช้ประกอบในการทำระบบแนะนำวัคซีนสำหรับเด็ก ซึ่งข้อมูลเหล่านี้มีความไม่ครบถ้วน ไม่สมบูรณ์ หรืออาจมีความซ้ำซ้อนกัน นอกจากนี้ ข้อมูลยังประกอบไปด้วยข้อมูลที่สามารถคำนวณได้ ซึ่งอยู่ในรูปของตัวเลข และข้อมูลที่ไม่สามารถคำนวณได้ เช่น ข้อมูลที่อยู่ในรูปข้อความ หรือข้อความปนตัวเลข หรือข้อมูลอาจมีความไม่สมบูรณ์เช่น ข้อมูลคอลัมน์ J, K, M, N และ O มีข้อมูล missing จำเป็นจะต้องมีการทำ Data Cleaning คือ การทำความสะอาดข้อมูล ซึ่งเป็นขั้นตอนสำหรับการคัดกรองข้อมูลที่เป็นส่วนรบกวน หรือข้อมูลที่ไม่เกี่ยวข้องออกไป ต้องมีการปรับปรุง หรือมีการแทนที่ เพื่อให้ข้อมูลมีคุณภาพ พร้อมนำไปใช้งานได้ ดังแสดงในภาพที่ 3-3 เป็นการเติมข้อมูลที่ขาดหายไปด้วยค่าเฉลี่ย ยกตัวอย่างเช่น พบว่ามีข้อมูลอายุหายไป 1 ตัว แต่มีข้อมูลอายุที่มี ได้แก่ 1, 3, 5 และ 1 ขั้นถัดไปจะทำการคำนวณหาค่าเฉลี่ย เมื่อคำนวณได้ผลลัพธ์ คือ 2.5 นำผลลัพธ์ที่ได้จากการคำนวณไปใส่ ณ ตำแหน่งที่ข้อมูลขาดหายไป

	age	gender	height	weight
1180	83	1	81.873519	11.253136
1181	83	1	81.873519	11.253136
1182	83	1	81.873519	11.253136
1183	84	1	109.900000	16.100000
1184	84	1	111.900000	16.900000
1185	84	1	114.700000	18.800000
1186	84	1	119.300000	21.000000
1187	84	1	124.200000	25.600000
1188	84	1	126.900000	27.200000
1189	84	1	128.600000	30.000000

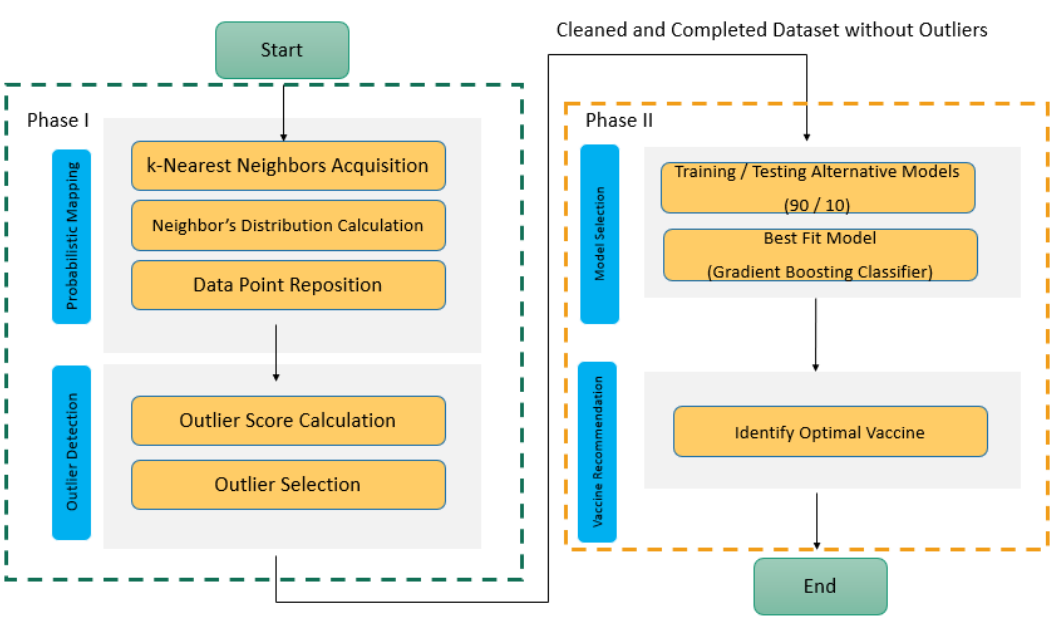
ภาพที่ 3-3 แสดงการเติมข้อมูลด้วยค่าเฉลี่ย

รวมถึงต้องมีการแปลงข้อมูลให้อยู่ในรูปที่สามารถวิเคราะห์ได้ อาทิ ข้อมูลรายชื่อวัคซีนซึ่งโดยทั่วไปแล้ว จะเป็นข้อมูลที่ประกอบด้วยตัวหนังสือผสมกับตัวเลข ซึ่งไม่สามารถนำไปคำนวณเพื่อการวิเคราะห์ได้ ต้องทำการแปลงข้อมูล ให้เป็นชนิดเดียวกัน หรืออาจมีการจัดกลุ่มข้อมูล ก่อนการนำไปใช้งานในเฟสต่อไป รวมไปถึงการตรวจหาและกำจัด Outliers โดยข้อมูลที่ส่งต่อไปยังเฟสที่ 2 จะเป็นชุดข้อมูลที่ผ่านการทำความสะอาดแล้ว มีการเติมข้อมูล และมีการกำจัดข้อมูลผิดปกติออกไป (cleaned and completed dataset without outliers)

หลังจากดำเนินการใน Phase I แล้ว ใน Phase II จะเป็นในส่วนของการศึกษาขั้นตอนวิธีเพื่อใช้ในการแนะนำวัคซีนสำหรับเด็ก จากข้อมูลที่ผ่านการจัดการ โดยการทำความสะอาดข้อมูลและเติมข้อมูล และการกำจัดข้อมูลที่ผิดปกติ ที่ดำเนินการได้แล้วใน Phase I ข้อมูลดังกล่าวจะมีความครบถ้วนและถูกต้อง เหมาะสำหรับนำไปใช้ในการวิเคราะห์เพื่อการแนะนำวัคซีน ที่เป็นเรื่องเดียวกันจะถูกจัดกลุ่มอย่างถูกต้อง ทำให้ได้ข้อมูลที่ครบถ้วนและเหมาะสมสำหรับการแนะนำวัคซีนรายคนสำหรับเด็กต่อไป

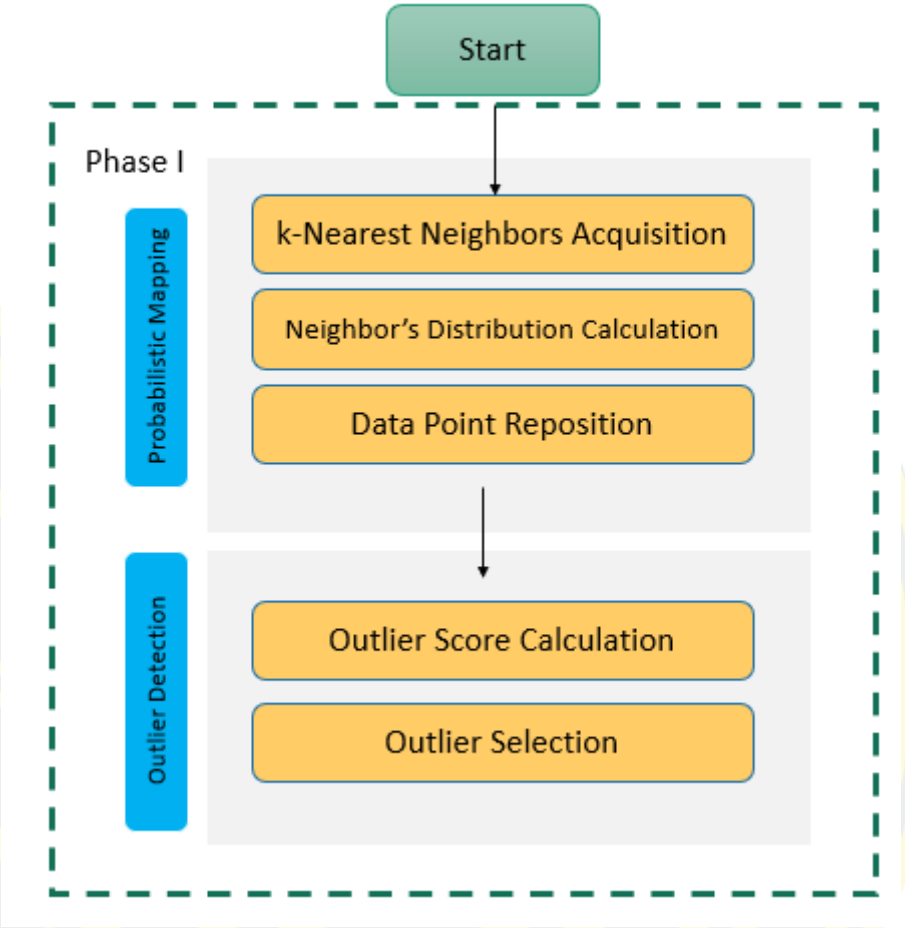
3.2 ขั้นตอนการดำเนินงานวิจัย

ในการดำเนินงานวิจัยตามกรอบการวิจัยที่ระบุในภาพ 3.1 จะประกอบไปด้วย 2 เฟส 4 ขั้นตอนหลัก และ 8 ขั้นตอนย่อย ดังแสดงในภาพ 3-3 ขั้นตอนการดำเนินงานใน Phase I และ Phase II



ภาพที่ 3-4 ขั้นตอนการดำเนินงานใน Phase I และ Phase II

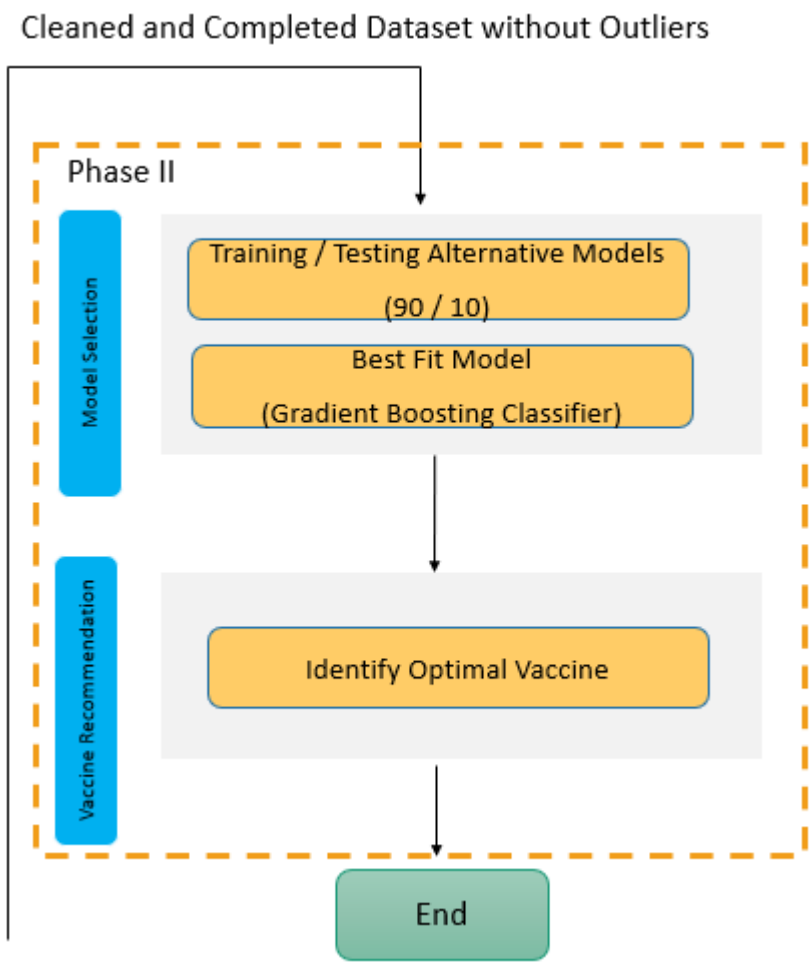
จากภาพที่ 3-3 ใน Phase I แบ่งเป็น 2 ขั้นตอนหลัก คือ 1) Probabilistic Mapping โดยแบ่งขั้นตอนเป็น 3 ขั้นตอนย่อย คือ (1.1) ขั้นตอน k-Nearest Neighbors Acquisition จะเป็นการเลือกจุดข้อมูลที่ใกล้ที่สุด (1.2) ขั้นตอน Neighbor's Distribution Calculation เป็นการคำนวณการกระจายความน่าจะเป็นของข้อมูล และ (1.3) ขั้นตอน Data Point Reposition จะทำการย้ายจากข้อมูลเดิมไปยังจุดที่มีความน่าจะเป็นสูงสุด 2) Outlier Detection มีขั้นตอนย่อย 2 ขั้นตอน คือ (2.1) ขั้นตอน Outlier Score Calculation จะเป็นการคำนวณหาค่า Score Outlier ส่วนในขั้นตอน (2.2) Outlier Selection เป็นการเลือกข้อมูลที่เป็น Outlier ดังแสดงในภาพ 3-5



ภาพที่ 3-5 ขั้นตอนการดำเนินงาน Phase I

ใน Phase II แบ่งเป็น 2 ขั้นตอนหลัก คือ 1) Model Selection โดยแบ่งเป็น 2 ขั้นตอนย่อย คือ (1.1) ขั้นตอน Training / Testing Alternative Models เป็นการทดลองใช้ขั้นตอนวิธีทั้งหมด 11 วิธี มาทดสอบ เพื่อเลือกโมเดลที่ดีที่สุด โดยทำการแบ่งข้อมูลในการ Training เป็น 90 และข้อมูลที่ใช้ Testing เป็น 10 จากนั้นในขั้นตอน (1.2) Best Fit Model เป็นการคำนวณหาค่าความถูกต้องของทุกขั้นตอนวิธีและพบว่า ขั้นตอนวิธี Gradient Boosting Classifier มีค่าความถูกต้องมากที่สุด 2) Vaccine Recommendation มี 1 ขั้นตอนย่อย คือ (2.1) ขั้นตอน Identify Optimal Vaccine ในขั้นตอนนี้จะเป็นการระบุวัคซีนสำหรับเด็กแต่ละราย โดยในขั้นตอนนี้ จะเป็นขั้นตอนศึกษาขั้นตอนวิธี เพื่อทำการเลือกขั้นตอนวิธีที่เหมาะสมกับข้อมูลคลินิกเด็กสุขภาพดี สำหรับในเลือกวัคซีนกับเด็กรายคน เพื่อประกอบการวินิจฉัยของบุคลากรทางการแพทย์ ดังแสดงในภาพ 3-6

โดยระบบจะนำข้อมูลที่พร้อมสำหรับการนำไปประมวลผล ไปทดสอบกับขั้นตอนวิธีทั้งหมด 11 วิธี แล้ววัดค่าความถูกต้องของขั้นตอนวิธี โดยมีมิติของข้อมูลที่ใช้ในการดำเนินการ ได้แก่ ข้อมูลเพศ อายุ น้ำหนัก ส่วนสูง pdx, dx0, dx1,dx2 และวัคซีนที่ได้รับ



ภาพที่ 3-6 ขั้นตอนดำเนินงาน Phase II

3.3 ข้อมูลและขั้นตอนวิธีที่ใช้ในการดำเนินงานวิจัย

3.3.1 Dataset และ Algorithm (Phase I)

ชุดข้อมูลที่นำมาใช้ในการทดลอง เพื่อเปรียบเทียบประสิทธิภาพของขั้นวิธีที่นำเสนอ ประกอบด้วย 4 ชุดข้อมูล ดังนี้

- Stamp มีข้อมูลทั้งหมด 340 แถว และมีข้อมูลผิดปกติ 9.10%
- Arrhythmia มีข้อมูลทั้งหมด 450 แถว และมีข้อมูลผิดปกติ 15.8%

- Pima มีข้อมูลทั้งหมด 768 แถว และมีข้อมูลผิดปกติ 34.9%
- Parkinson มีข้อมูลทั้งหมด 195 แถว และมีข้อมูลผิดปกติ 75.4%

Probabilistic Mapping

Probabilistic Mapping เป็นการย้ายข้อมูลที่ได้จากการคำนวณความน่าจะเป็น ในขั้นตอนนี้ประกอบด้วย 3 ขั้นตอน ในขั้นตอนที่ 1 K-Nearest Neighbors Acquisition ในขั้นตอนที่ 2 Neighbor's Distribution Calculation และในขั้นตอนที่ 3 Data Point Reposition อธิบายได้ดังนี้

K-Nearest Neighbors Acquisition

K-Nearest Neighbors Acquisition จะเป็นการเลือกจุดข้อมูลที่ใกล้ที่สุด โดยใช้ขั้นตอนวิธี k-Nearest Neighbors มีหลักการทำงาน ดังนี้

ขั้นแรก : กำหนดขนาดของ K

ขั้นสอง : คำนวณระยะห่างของข้อมูลที่ต้องพิจารณากับกลุ่มข้อมูล

ขั้นสาม : จัดเรียงลำดับของระยะห่าง และเลือกพิจารณาชุดข้อมูลที่ใกล้กับจุดที่สนใจ พิจารณาจากตามจำนวน K ที่กำหนดไว้

ขั้นสี่ : พิจารณาข้อมูลจำนวน K ชุด และสังเกตว่าจุดไหนใกล้กับจุดที่สนใจมากที่สุด

Neighbor's Distribution Calculation

Neighbor's Distribution Calculation จะเป็นการคำนวณการกระจายความน่าจะเป็นของข้อมูล หลังจากที่หาข้อมูลเท่ากับจำนวน K ที่กำหนดแล้ว ข้อมูลแต่ละชุดที่ทำการหาระยะห่างที่ใกล้ที่สุดจะถูกนำไปคำนวณหาค่าการกระจายตัวของข้อมูล ด้วยขั้นตอน Truncated Gaussian Distribution (TGD) ดังสมการ (3-1)

$$(x;\mu,\sigma,a,b) = \frac{1}{\sigma}\frac{\Phi(\frac{x-\mu}{\sigma})}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi(\frac{a-\mu}{\sigma})} \qquad (3.1)$$

โดยให้

x คือ การแจกแจงแบบปกติที่มีค่าเฉลี่ย μ และความแปรปรวน σ^2 ภายในช่วง (a,b) จากนั้นเงื่อนไข X บน $a \leq x \leq b$ มีการแจกแจงแบบเกาส์เซียนที่ถูกตัดทอน (Truncated Gaussian Distribution)

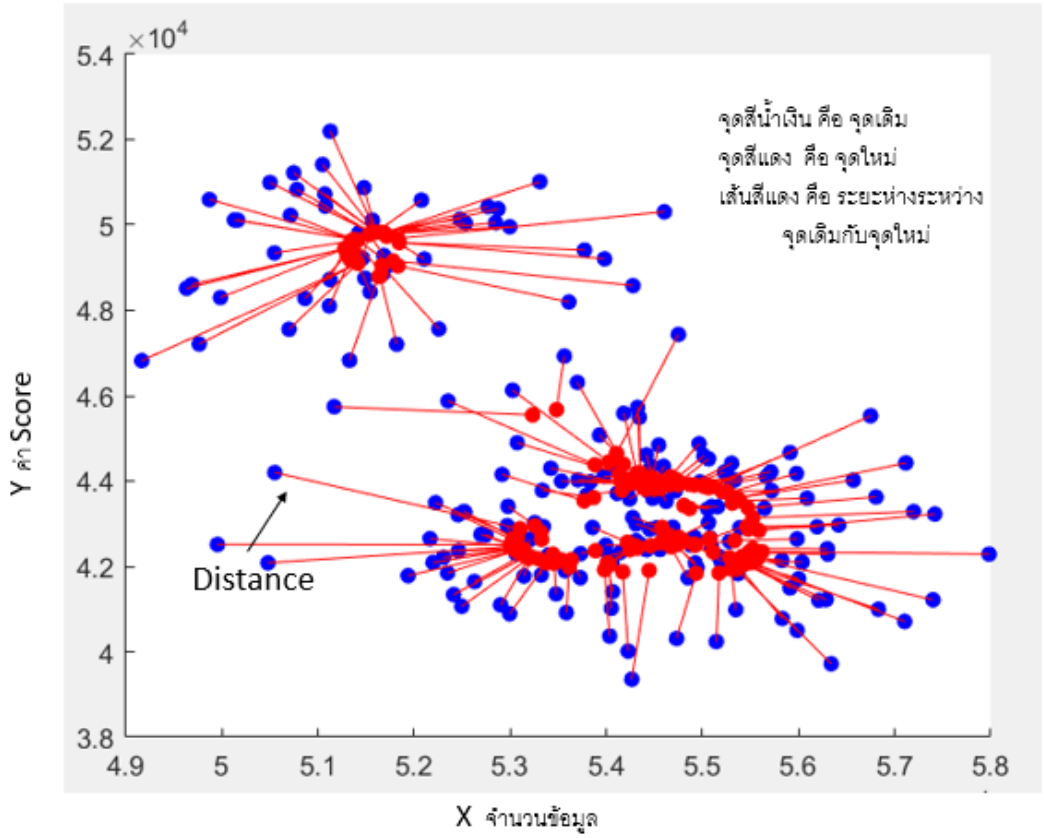
Data Point Reposition

Data Point Reposition เป็นการย้ายข้อมูลเดิมไปยังจุดข้อมูลที่มีความน่าจะเป็นสูงสุด โดยมีขั้นตอน ดังนี้

ขั้นตอนแรก : เลือกจุดเริ่มต้นจากชุด *K-Nearest Neighbors (k-NN)*

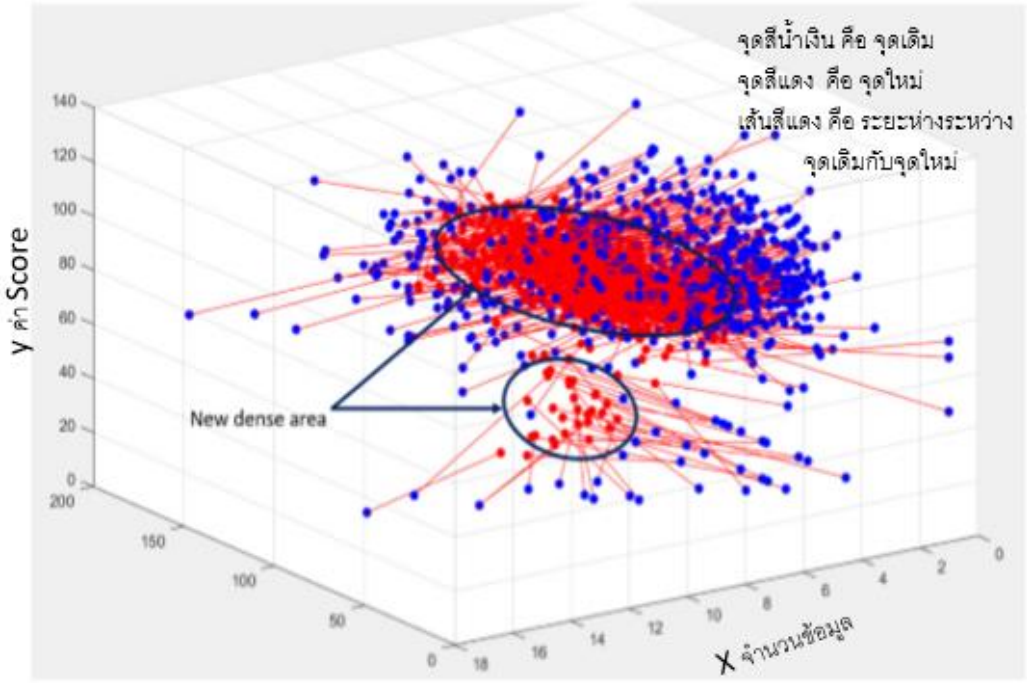
ขั้นตอนสอง : จุดแต่ละจุดคำนวณหาค่าความน่าจะเป็นของการเป็นสมาชิกในแต่ละกลุ่ม

ขั้นตอนสาม : เลือกน้ำหนักที่มีค่าความน่าจะเป็นสูงที่สุด ดังภาพที่ *3-7*



ภาพที่ 3-7 แสดงภาพจุดเดิมไปยังจุดใหม่

จากภาพที่ 3-7 แสดงการย้ายจุดข้อมูลไปยังตำแหน่งใหม่ ใช้การคำนวณค่าความน่าจะเป็นสูงสุด ใช้วิธี Truncated Gaussian Distribution โดยที่ จุดสีน้ำเงิน คือ จุดเดิม ส่วนจุดสีแดง คือ จุดใหม่ และเส้นสีแดง คือ ระยะห่างระหว่างจุดสีน้ำเงินและสีแดง



ภาพที่ 3-8 ผลลัพธ์พื้นที่หนาแน่นหลังการทำ ***Mapping***

จากภาพที่ 3-8 แสดงการย้ายจุดข้อมูลไปยังตำแหน่งใหม่ ใช้การคำนวณค่าความน่าจะเป็นสูงสุด ใช้วิธี Truncated Gaussian Distribution โดยที่ จุดสีน้ำเงิน คือ จุดเดิม ส่วนจุดสีแดง คือ จุดใหม่ และเส้นสีแดง คือ ระยะห่างระหว่างจุดเดิมกับจุดใหม่

Outlier Detection

Outlier Detection เป็นการตรวจหาค่าผิดปกติในข้อมูลปกติ มีขั้นตอน 2 ขั้นตอน ได้แก่ ในขั้นตอนที่ 1 Outlier Score Calculation และในขั้นตอนที่ 2 Outlier Selection อธิบายได้ ดังนี้

Outlier Score Calculation

Outlier Score Calculation ในขั้นตอนนี้จะเป็นการคำนวณหา Score Outlier ข้อมูลที่ได้จะมี 2 ชุด คือ ข้อมูลเดิม และข้อมูลใหม่ ซึ่งจะใช้ระยะทางคำนวณหาค่า Score ของ Outlier สามารถคำนวณหาค่า Score Outlier ดังภาพที่ 3-9 และภาพที่ 3-10 ดังสมการ (3.4)

$$S = \sum_{i=1}^{n} \left| X_i - X_{map(i)} \right| + \frac{\sum (X_i - \mu)^2}{n-1} \qquad (3.4)$$

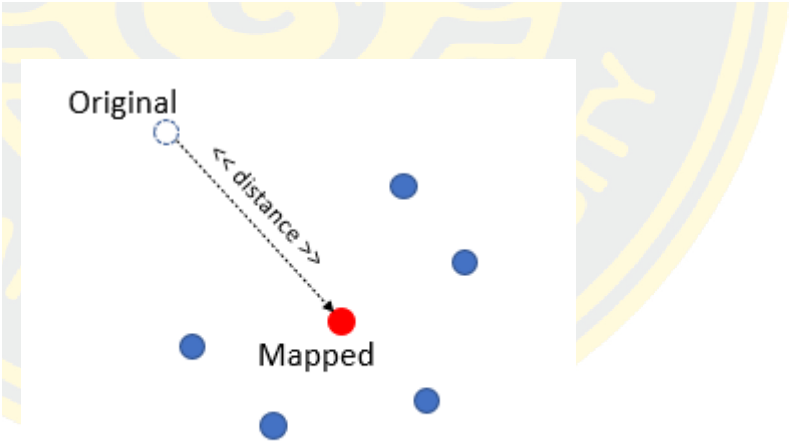
โดยให้

S เป็นคะแนนผิดปกติ

X_i คือ ข้อมูลจากชุดดั้งเดิม

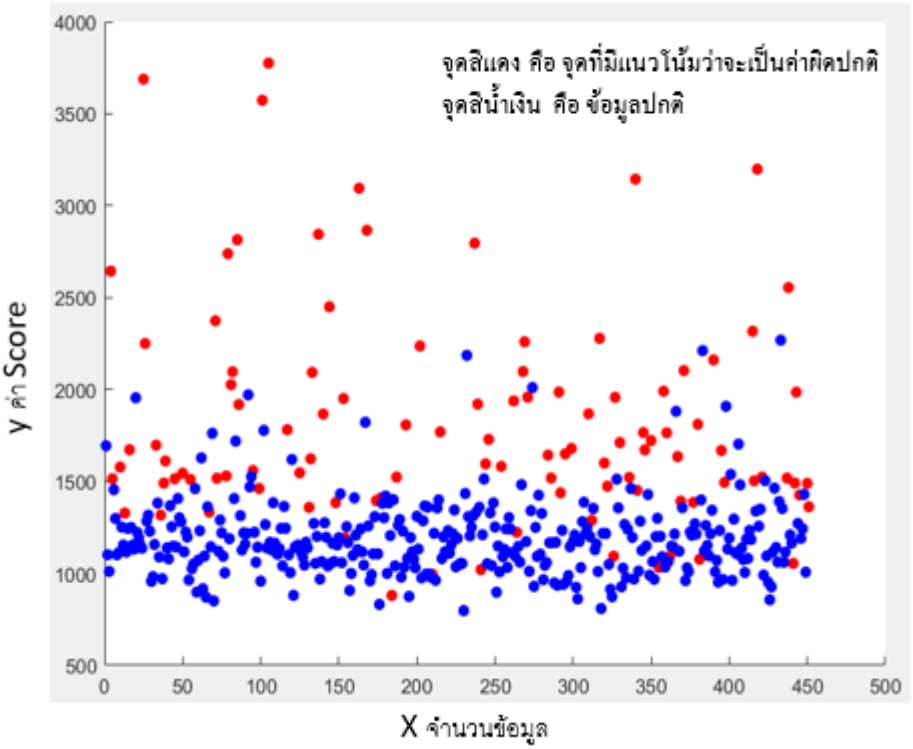
$X_{map(i)}$ คือ ข้อมูลจากอินสแตนซ์ที่แมป

μ คือ ค่าเฉลี่ยของเซต k-NN คะแนนจะใช้ในขั้นตอนการเลือกค่าผิดปกติ



ภาพที่ 3- 9 ตัวอย่างการขยับระยะทางระหว่างข้อมูลที่แมปกับข้อมูลเดิม

จากภาพที่ 3-9 แสดงตัวอย่างการขยับระยะทางระหว่างจุดข้อมูลเดิมกับจุดข้อมูลที่ขยับ โดยที่จุดสีน้ำเงิน คือ จุดข้อมูลเดิม และสีแดง คือ จุดที่ขยับ แล้วทำการคำนวณหาค่า Score Outlier



ภาพที่ 3-10 แสดงข้อมูลที่ได้จากการคำนวณ Outlier Score

จากภาพที่ 3-10 เป็นข้อมูลที่ได้จากการคำนวณ Outlier Score โดยจุดสีน้ำเงิน คือ จุดข้อมูลปกติ และจุดสีแดง คือ จุดที่มีแนวโน้มว่าจะเป็นค่าผิดปกติ

Outlier Selection

Outlier Selection จะเป็นการเลือกข้อมูลที่เป็น Outlier หลังจากคำนวณ Outlier Score ของทุกจุดข้อมูลแล้ว จากนั้นทำการเรียง Outlier Score จากมากไปน้อย ในการเลือกค่า Outlier จากชุดข้อมูลทั้งหมด ได้เสนอวิธีการเพื่อกำหนดคะแนนค่าผิดปกติ โดยใช้วิธี Top-N สามารถกำหนดจำนวนจุดที่เลือกได้ ดังสมการ (3-5)

$$Top_n = \left\lfloor \frac{1}{2}(R_{out}N \right\rfloor, Top_n \epsilon I^+ \qquad (3.5)$$

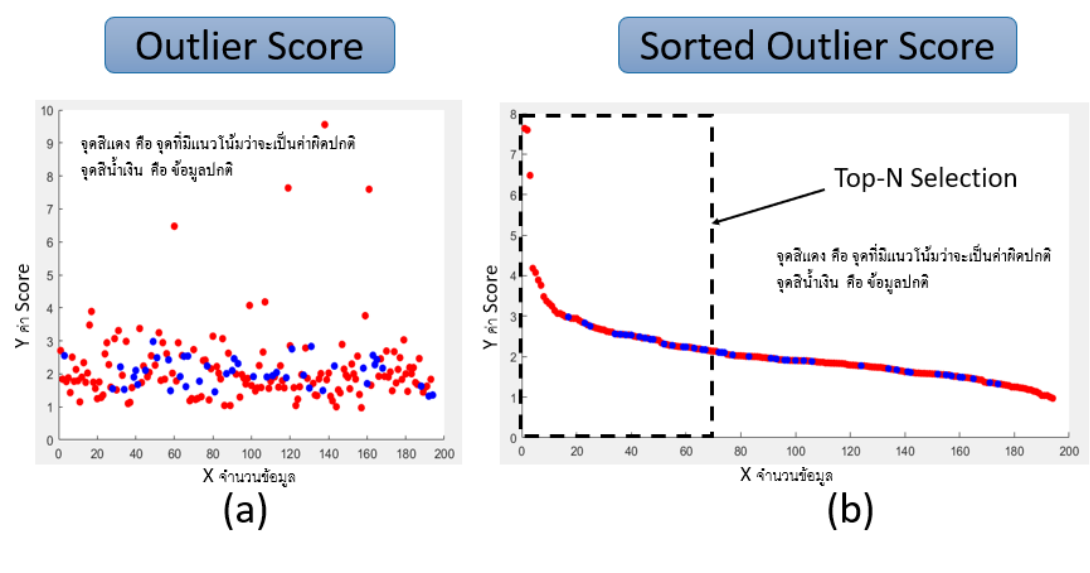
โดยที่

Top_n คือ จำนวนคะแนน Top-N outlier

R_{out} คือ อัตราส่วนหรือเปอร์เซ็นต์ของชุดข้อมูลที่ผิดปกติ

N คือ จำนวนอินสแตนซ์ทั้งหมด

ตัวอย่างเช่น ชุดข้อมูล CWC มีค่าผิดปกติ *4%* ชุดข้อมูล *Top-N* ของ *CWC* สามารถคำนวณได้ $\frac{1}{2}(0.04)(195) \approx 39$ จุด



ภาพที่ 3-11 แสดงขั้นตอนการคำนวณหาค่า Outlier Score และ Sorted Outlier Score

จากภาพที่ *3-11* ในภาพที่ *3-11 (a)* เป็นขั้นตอนคำนวณหาค่า *Outlier Score* โดยที่จุดสีแดง คือ จุดที่มีแนวโน้มว่าจะเป็นค่าผิดปกติและจุดสีน้ำเงิน คือ ข้อมูลปกติ ในขั้นตอนนี้ คะแนนผิดปกติที่คำนวณยังไม่มีการเรียงลำดับ ภาพที่ *3-11 (b)* จะเป็นขั้นตอน *Sorted Outlier Score* ในการเลือกข้อมูลจะต้องเรียงลำดับจากมากไปหาน้อย แล้วทำการใช้วิธี *Top-N* ในการเลือกข้อมูลผิดปกติ

3.3.2 Dataset และ Algorithm (Phase II)

ชุดข้อมูลที่นำมาใช้ในการทดลอง เพื่อเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอประกอบด้วย *2* ชุดข้อมูล ดังนี้

- Wine มีข้อมูลทั้งหมด 1599 แถว มีมิติข้อมูล 12 มิติ และมีคลาส 6 คลาส

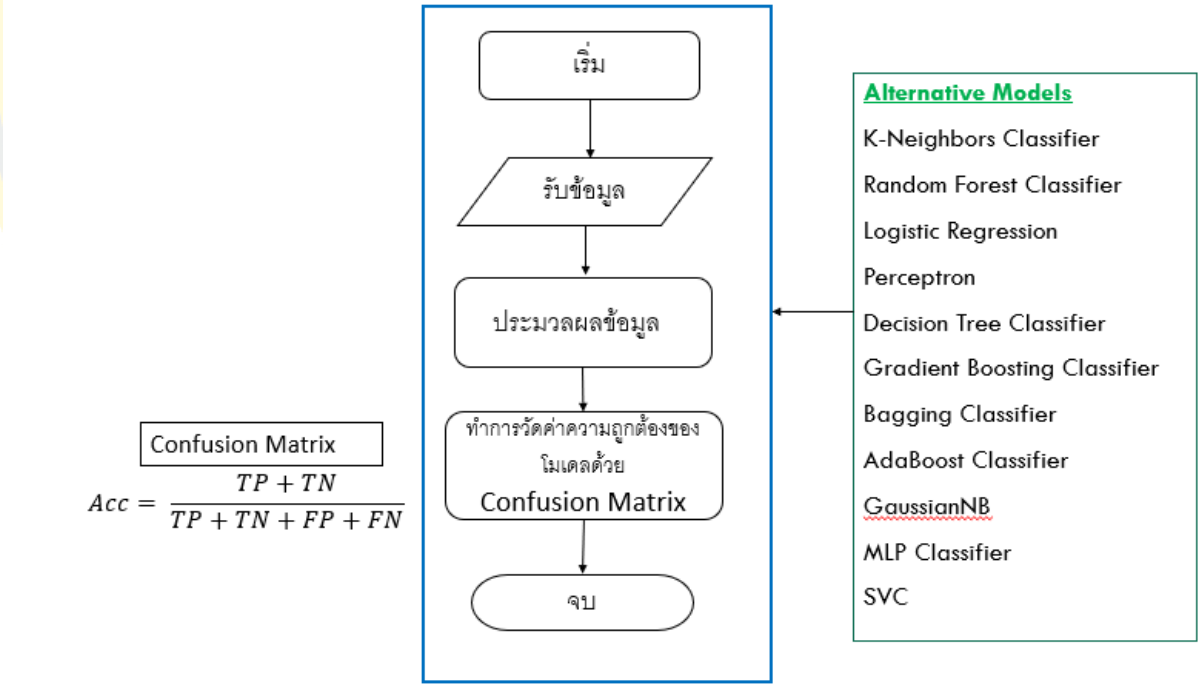
- Iris มีข้อมูลทั้งหมด 150 แถว มีมิติข้อมูล 6 มิติ และมีคลาส 3 คลาส

Model Selection

Model Selection เป็นขั้นตอนการศึกษาขั้นตอนวิธี 11 วิธี เพื่อนำมาเปรียบเทียบขั้นตอนวิธีที่ให้ค่าความถูกต้องมากที่สุด โดยแบ่ง 2 ขั้นตอน ได้แก่ ขั้นตอน Training /Testing Alternative Models และขั้นตอน Best Fit Model อธิบายได้ ดังนี้

Training / Testing Alternative Models

Training / Testing Alternative Models เป็นการทดลองใช้ขั้นตอนวิธีทั้งหมด 11 วิธี มาทดสอบ เพื่อเลือกโมเดลที่ดีที่สุด โดยทำการแบ่งข้อมูล Training คือ 90 และ Testing คือ 10 แล้วทำการประมวลผลด้วยขั้นตอนวิธีทั้งหมด 11 วิธี ดังภาพที่ 3-12



ภาพที่ 3-12 แสดงขั้นตอนในการทดลองขั้นตอนวิธีทั้งหมด 11 วิธี

จากภาพที่ 3-9 แสดงขั้นตอนในการทดลองขั้นตอนวิธีทั้งหมด 11 วิธี จากนั้นทำการประมวลผลข้อมูล แล้วทำการวัดค่าความถูกต้องของขั้นตอนวิธีแต่ละวิธีด้วย Confusion Matrix

Best Fit Model

Best Fit Model จะเป็นการคำนวณหาค่าความถูกต้องของทุกขั้นตอนวิธีทั้ง 11 วิธี ได้แก่ K-Neighbors Classifier, Random Forest Classifier, Logistic Regression, Perceptron, Decision Tree Classifier, Gradient Boosting Classifier, Bagging Classifier, AdaBoost Classifier, GaussianNB, MLP Classifier และ SVC เมื่อทำการวัดค่าความถูกต้องของโมเดลด้วย Confusion Matrix แล้ว พบว่าขั้นตอนวิธีการ Gradient Boosting Classifier ให้ค่าความถูกต้องมากที่สุด

บทที่ 4

ผลการทดลอง

ในบทนี้จะกล่าวถึง ผลการทดลอง ซึ่งจากวิธีการที่นำเสนอในบทที่ 3 ของงานวิจัยนี้ ซึ่งแบ่งออกเป็น 2 เฟส (Phase) คือ เฟสที่ 1 เป็นการดำเนินการเพื่อจัดการกับข้อมูลที่ไม่สมบูรณ์และข้อมูลผิดปกติ และเฟสที่ 2 เป็นการดำเนินการเพื่อหาขั้นตอนวิธีที่เหมาะสมในการระบุวัคซีนรายคนจากประวัติของเด็ก ดังนี้

4.1 **ข้อมูลที่ใช้ในการทดลอง (data sets)**

ข้อมูลที่ใช้ในการทดลองของงานวิจัยนี้ ประกอบไปด้วยข้อมูลดังนี้

4.1.1 **ข้อมูลเซตที่ 1**

เซตที่ 1 คือ ข้อมูลที่ใช้ในการทดลองเฟสที่ 1 เพื่อจัดการกับข้อมูลที่ไม่สมบูรณ์และข้อมูลผิดปกติ โดยจะประกอบไปด้วยข้อมูลสองส่วน คือ ส่วนที่ 1 เป็นข้อมูลจากฐานข้อมูลเด็กสุขภาพดี หรือ CWC ซึ่งเป็นฐานข้อมูลเป้าหมายของงานวิจัย และส่วนที่ 2 ข้อมูลที่เผยแพร่สาธารณะ เพื่อใช้ในการทดสอบและเปรียบเทียบผล ได้แก่ Arrhythmia, Pima, Parkinson, และ Stamps ดังมีตัวอย่างข้อมูล ดังนี้

Age	Sex	height	weight	QRS	P-R	Q-T	T interval	P interval	QRS
56	1	165	64	81	174	401	149	39	25
54	0	172	95	138	163	386	185	102	96
55	0	175	94	100	202	380	179	143	28
75	0	190	80	88	181	360	177	103	-16
13	0	169	51	100	167	321	174	91	107
40	1	160	52	77	129	377	133	77	77
49	1	162	54	78	0	376	157	70	67
44	0	168	56	84	118	354	160	63	61
50	1	167	67	89	130	383	156	73	85
62	0	170	72	102	135	401	156	83	72
45	1	165	86	77	143	373	150	65	12
54	1	172	58	78	155	382	163	81	-24
30	0	170	73	91	180	355	157	104	68
44	1	160	88	77	158	399	163	94	46
47	1	150	48	75	132	350	169	65	36
47	0	171	59	82	145	347	169	61	77

ตาราง 7 ตัวอย่างข้อมูล Arrhythmia

Prenancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0

ตาราง 8 ตัวอย่างข้อมูล Pima

MDVP: Fo(Hz)	MDVP: Fhi(Hz)	MDVP: Flo(Hz)	MDVP: Jitter(%)	MDVP: Jitter(Abs)	MDVP: RAP	MDVP: PPQ	Jitter: DDP	MDVP: Shimmer	MDVP: Shimmer(dB)
119.992	157.302	74.997	0.00784	0.00007	0.0037	0.00554	0.01109	0.04374	0.426
122.4	148.65	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134	0.626
116.682	131.111	111.555	0.0105	0.00009	0.00544	0.00781	0.01633	0.05233	0.482
116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492	0.517
116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425	0.584
120.552	131.162	113.787	0.00968	0.00008	0.00463	0.0075	0.01388	0.04701	0.456
120.267	137.244	114.82	0.00333	0.00003	0.00155	0.00202	0.00466	0.01608	0.14
107.332	113.84	104.315	0.0029	0.00003	0.00144	0.00182	0.00431	0.01567	0.134
95.73	132.068	91.754	0.00551	0.00006	0.00293	0.00332	0.0088	0.02093	0.191
95.056	120.103	91.226	0.00532	0.00006	0.00268	0.00332	0.00803	0.02838	0.255
88.333	112.24	84.072	0.00505	0.00006	0.00254	0.0033	0.00763	0.02143	0.197
91.904	115.871	86.292	0.0054	0.00006	0.00281	0.00336	0.00844	0.02752	0.249
136.926	159.866	131.276	0.00293	0.00002	0.00118	0.00153	0.00355	0.01259	0.112
139.173	179.139	76.556	0.0039	0.00003	0.00165	0.00208	0.00496	0.01642	0.154
152.845	163.305	75.836	0.00294	0.00002	0.00121	0.00149	0.00364	0.01828	0.158
142.167	217.455	83.159	0.00369	0.00003	0.00157	0.00203	0.00471	0.01503	0.126
144.188	349.259	82.764	0.00544	0.00004	0.00211	0.00292	0.00632	0.02047	0.192
168.778	232.181	75.603	0.00718	0.00004	0.00284	0.00387	0.00853	0.03327	0.348
153.046	175.829	68.623	0.00742	0.00005	0.00364	0.00432	0.01092	0.05517	0.542

ตาราง 9 ตัวอย่างข้อมูล Parkinson

Height	Lenght	Area	Eccen	P_ black	P_ and	Mean_TR	Black pix	Black and	WB_ trans	Outcome
5	7	35	1.4	0.4	0.657	2.33	14	23	6	1
6	7	42	1.167	0.429	0.881	3.6	18	37	5	1
6	18	108	3	0.287	0.741	4.43	31	80	7	1
5	7	35	1.4	0.371	0.743	4.33	13	26	3	1
6	3	18	0.5	0.5	0.944	2.25	9	17	4	1
5	8	40	1.6	0.55	1	2.44	22	40	9	1
6	4	24	0.667	0.417	0.708	2.5	10	17	4	1
5	6	30	1.2	0.333	0.333	10	10	10	1	1
5	5	25	1	0.4	0.52	10	10	13	1	1
5	7	35	1.4	0.486	0.914	8.5	17	32	2	1
5	2	10	0.4	0.8	1	8	8	10	1	1
5	3	15	0.6	0.533	0.733	8	8	11	1	1
5	6	30	1.2	0.433	0.733	13	13	22	1	1
5	7	35	1.4	0.429	0.857	5	15	30	3	1
6	7	42	1.167	0.405	0.881	4.25	17	37	4	1
5	8	40	1.6	0.375	0.475	15	15	19	1	1
6	7	42	1.167	0.405	0.952	4.25	17	40	4	1
5	19	95	3.8	0.232	0.4	5.5	22	38	4	1
5	5	25	1	0.4	0.56	5	10	14	2	1
4	6	24	1.5	0.417	0.5	10	10	12	1	1

ตาราง 10 ตัวอย่างข้อมูล Stamps

เพศ	age_y	น้ำหนัก (กก.)	ส่วนสูง (ซม.)	pdx	dx0	dx1	dx2	epi_vaccine_name (วัคซีนที่ได้รับ)
0	5	18.00	89	43	44	45	46	7
0	1	9.80	80	55	43	65	58	2
0	1	9.80	80	55	43	65	58	1
0	4	16.60	133	55	81	43	58	3
0	4	16.60	133	55	81	43	58	4
0	2	9.80	79	55	84	43	58	2
0	2	9.80	79	55	84	43	58	1
1	2	10.00	80	55	21	43	58	2
1	2	10.00	80	55	21	43	58	1
0	1	13.00	83	55	33	43	58	2
0	1	13.00	83	55	33	43	58	1
1	4	23.80	104	55	3	43	58	3
1	4	23.80	104	55	3	43	58	4
1	4	15.10	98	55	32	43	58	3
1	4	15.10	98	55	32	43	58	4
1	4	13.70	101	55	33	43	58	3
1	4	13.70	101	55	33	43	58	4
0	2	11.00	81	55	21	32	58	2
0	2	11.00	81	55	21	32	58	1

ตาราง 11ตัวอย่างข้อมูลจากฐานคลินิกเด็กสุขภาพดี (CWC)

4.1.2 ข้อมูลเซตที่ 2

เซตที่ 2 คือ ข้อมูลที่ใช้ในการทดลองเฟสที่ 2 เพื่อหาขั้นตอนวิธีที่เหมาะสมในการระบุวัคซีนรายคนจากประวัติคนไข้ของเด็ก โดยจะประกอบไปด้วยข้อมูลสองส่วน คือ ส่วนที่ 1 เป็นข้อมูลจากฐานข้อมูลเด็กสุขภาพดี หรือ CWC ซึ่งเป็นฐานข้อมูลเป้าหมายของงานวิจัย และส่วนที่ 2 ข้อมูลที่เผยแพร่สาธารณะ เพื่อใช้ในการทดสอบและเปรียบเทียบผล ได้แก่ Wine และ Iris ดังมีตัวอย่างข้อมูลดังนี้

fixed acidity	volatile acidity	citric acid	Residual sugar	Chlorides	free sulfur dioxide	total sulfur dioxide	Density	pH	Sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4

ตาราง 12 ตัวอย่างข้อมูล Wine

Id	Sepal LengthCm	Sepal WidthCm	Petal LengthCm	Petal WidthCm	Species
1	5.1	3.5	1.4	0.2	1
2	4.9	3	1.4	0.2	1
3	4.7	3.2	1.3	0.2	1
4	4.6	3.1	1.5	0.2	1
5	5	3.6	1.4	0.2	1
6	5.4	3.9	1.7	0.4	1
7	4.6	3.4	1.4	0.3	1
8	5	3.4	1.5	0.2	1
9	4.4	2.9	1.4	0.2	1
10	4.9	3.1	1.5	0.1	1
11	5.4	3.7	1.5	0.2	1
12	4.8	3.4	1.6	0.2	1
13	4.8	3	1.4	0.1	1
14	4.3	3	1.1	0.1	1
15	5.8	4	1.2	0.2	1
16	5.7	4.4	1.5	0.4	1

17	5.4	3.9	1.3	0.4	1
18	5.1	3.5	1.4	0.3	1
19	5.7	3.8	1.7	0.3	1

ตาราง 13ตัวอย่างข้อมูล Iris

4.2 ผลการทดลอง

4.2.1 ผลการทดลองสำหรับเฟสที่ 1

ในการจัดการข้อมูลผิดปกติ ผู้วิจัยได้นำเสนอขั้นตอนวิธี Probabilistic Mapped Mean Shift (PMMS) เพื่อตรวจจับข้อมูลที่มีความผิดปกติ (outlier) เพื่อกำจัดข้อมูลเหล่านั้น โดยก่อนนำข้อมูลมาตรวจหาความผิดปกติของข้อมูลนั้น จะมีการจัดการกับข้อมูลที่ไม่สมบูรณ์ (incomplete) ด้วยการเติมข้อมูลที่ขาดหายไป โดยใช้ค่าเฉลี่ยของข้อมูล (average value)ที่มีพฤติกรรมใกล้เคียงกัน

โดยขั้นตอนวิธีที่นำเสนอ คือ PMMS ได้รับการทดสอบความมีประสิทธิภาพของขั้นตอนวิธีโดยการทดสอบเปรียบเทียบกับขั้นตอนวิธีอื่น ๆ คือ DOD+, MOD+, DOD และ MOD และใช้ข้อมูลเหล่านี้ในการทดสอบ คือ ข้อมูล CWC (ข้อมูลจากคลินิกเด็กสุขภาพดี) เทียบกับข้อมูลที่เผยแพร่สาธารณะ คือ ข้อมูล Stamps ข้อมูล Arrhythmia ข้อมูล Pima และข้อมูล Parkinson และใช้ตัววัดผล คือ Confusion Matrix ดังสมการ 4.1 ในการวัดผลค่าความถูกต้องของแต่ละขั้นตอนวิธีและได้ผลลัพธ์ ดังตารางที่ 14

$$Acc = \frac{TP}{TP+FP} \quad (4.1)$$

Dataset (Outliers)	CWC. (4%)	Stamps (9.1%)	Arrh (15%)	Pima (34.9%)	Parkinson (75.4%)	Average
PMMS (proposed)	**0.93**	**0.94**	**0.80**	**0.75**	**0.72**	**0.80**
DOD+	0.91	0.89	0.73	0.71	0.63	0.74
MOD+	0.89	0.91	0.72	0.73	0.66	0.76
DOD	0.92	0.81	0.73	0.68	0.68	0.73
MOD	0.92	0.78	0.74	0.65	0.60	0.69
LOF	-	0.53	0.73	0.60	0.56	0.61
ODIN	-	0.58	0.70	0.56	0.45	0.57
NC	-	0.50	0.60	0.52	0.46	0.52
KNN	-	0.89	0.74	0.72	0.54	0.72
ABOD	-	0.81	0.72	0.70	0.64	0.72

ตาราง 14แสดงผลการวัดโดยใช้ Confusion Matrix

4.2.2 ผลการทดลองสำหรับเฟสที่ 2

ในการดำเนินการศึกษาเพื่อเลือกขั้นตอนวิธีที่เหมาะสม สำหรับฐานข้อมูลเด็กสุขภาพดีนั้น ผู้วิจัย ได้นำขั้นตอนวิธีที่นิยมใช้จำนวน 11 ขั้นตอนวิธี คือ K-Neighbors Classifier, Random Forest Classifier, Logistic Regression, Perceptron, Decision Tree Classifier, Gradient Boosting Classifier, Bagging Classifier, AdaBoost Classifier, GaussianNB, MLP Classifier และ SVC โดยนำมาทดสอบกับฐานข้อมูลเด็กสุขภาพดี (CWC) ซึ่งมีจำนวน Class เท่ากับ 21 Class และมีมิติข้อมูล 9 มิติ และฐานข้อมูลสาธารณะอีก 2 ชุด คือ 1) ฐานข้อมูล Winequality-red มีจำนวน Class เท่ากับ 6 Class และมีมิติข้อมูล 12 มิติ และ 2) ฐานข้อมูล Iris มีจำนวน Class เท่ากับ 3 Class และมีมิติข้อมูล 6 มิติ และทำการวัดผลค่าความถูกต้องด้วย Confusion Matrix ได้ผลลัพธ์ ดังตาราง 15

Models	**Accuracy**		
	CWC (21 class)	Wine (6 class)	Iris (3 class)
K-Neighbors Classifier	0.11	0.44	0.94
Random Forest Classifier	0.49	**0.57**	0.98
Logistic Regression	0.17	**0.57**	0.97
Perceptron	0.11	0.41	0.66
Decision Tree Classifier	0.45	0.46	0.95
Gradient Boosting Classifier	**0.53**	0.56	0.96
Bagging Classifier	0.49	0.54	0.96
AdaBoost Classifier	0.31	0.53	0.96
GaussianNB	0.17	0.54	**0.99**
MLP Classifier	0.11	0.54	0.81
SVC	0.17	0.50	0.94

ตาราง 15แสดงผลการวัดโดยใช้ Accuracy mean

บทที่ 5

สรุปผลการวิจัย

5.1 สรุปผลการวิจัย

ในงานวิจัยนี้ ได้นำเสนอระบบแนะนำวัคซีนสำหรับคลินิกเด็กสุขภาพดี ซึ่งกรอบการทำงานของ แบ่งเป็น 2 เฟส คือ เฟสที่ 1 จะเป็นการจัดการข้อมูล โดยมีการทำความสะอาดข้อมูลและเติมข้อมูลด้วยค่าเฉลี่ย มีการกำจัดข้อมูลที่ผิด (outliers) ด้วยขั้นตอนวิธี Probabilistic Mapped Mean-Shift (PMMS) ซึ่งมีค่าความถูกเท่ากับ 93%, 94%, 80%, 75%, และ 72% เมื่อนำไปทดลองกับข้อมูล CWC, Stamps, Arrh, Pima และ Pakinson ตามลำดับ โดยค่าความถูกต้องที่ได้ดังกล่าวนี้ เป็นค่าความถูกต้องสูงสุด เมื่อเทียบกับขั้นตอนวิธีอื่น ที่ใช้ในการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอ เมื่อจัดการกับข้อมูลในเฟสที่ 1 เรียบร้อยแล้ว ทำให้ได้ชุดข้อมูล CWC ที่ผ่านการทำความสะอาดและมีการเติมข้อมูล รวมถึงได้มีการกำจัดข้อมูลผิดปกติออกไป (cleaned and completed dataset without outliers) ก็จะนำข้อมูล CWC ไปทดลองเพื่อหาขั้นตอนวิธีที่เหมาะสมในการแนะนำวัคซีนรายคนสำหรับเด็ก ในเฟสที่ 2 ซึ่งพบว่า ขั้นตอนวิธี Gradient Boosting Classifier ให้ค่าความถูกต้องสูงสุดอยู่ที่ 53% ซึ่งเป็นค่าที่สูงสุดจากขั้นตอนวิธีทั้งหมด 11 วิธี

5.2 ประโยชน์ที่ได้รับ

- ได้ขั้นตอนวิธีสำหรับการตรวจจับข้อมูลผิดปกติ โดยใช้วิธีการ Machine Learning ที่เหมาะสมกับลักษณะข้อมูลในฐานข้อมูลคลินิกเด็ก เพื่อเตรียมข้อมูลสำหรับนำไปใช้ในระบบแนะนำวัคซีน
- ได้ขั้นตอนวิธี ในการระบุชนิดของวัคซีนที่สอดคล้องกับข้อมูลรายคน สำหรับการแนะนำวัคซีนที่เหมาะสมกับช่วงอายุ พัฒนาการทางร่างกายของเด็กแต่ละคน เพื่อสนับสนุนการวินิจฉัยของแพทย์

5.3 **งานที่จะทำในอนาคต**

จากผลการทดลองในบทที่ 4 และข้อสังเกตจากผลการทดลองที่ได้ พบว่า ขั้นตอนวิธี Gradient Boosting Classifier ที่ให้ค่าความถูกต้องสูงสุดนั้น ในกระบวนการทำงานของขั้นตอนวิธีนี้ได้มีการใช้หลักการของ Ensemble Learning ในการดำเนินการเพื่อหาค่าคำตอบ ในขณะที่วิธีอื่นที่ใช้ในเปรียบเทียบประสิทธิภาพในการแนะนำวัคซีนรายคน ไม่มีการใช้หลักการดังกล่าว จึงอาจเป็นไปได้ว่า วิธีการแบบ Ensemble Learning มีความเหมาะสมกับข้อมูลที่มีจำนวนกลุ่มข้อมูลและมิติข้อมูลสูง ดังเช่นปรากฏในข้อมูลจากคลินิกเด็กสุขภาพดี

นอกจากการศึกษาความเป็นไปได้ของหลักการ Ensemble Learning ที่เหมาะสมกับข้อมูลที่มีจำนวนคลาสและมิติข้อมูลสูง ๆ แล้ว ยังต้องมีการศึกษาเพื่อหาวิธีการที่สามารถเพิ่มประสิทธิภาพให้โมเดลมีค่าความถูกต้องมากยิ่งขึ้นในการแนะนำวัคซีนรายบุคคล เพื่อให้ได้ขั้นตอนวิธีที่เหมาะสมมากยิ่งขึ้น ที่ทำให้การนำไปใช้งานจริงมีความน่าเชื่อถือ

ภาคผนวก

Outlier Detection in Wellness Data using Probabilistic Mapped Mean-Shift Algorithms

Siriwan Phongsasiri[1] and Suwanna Rasmequan[2]

ABSTRACT: In this paper, the Probabilistic Mapped Mean-Shift Algorithm is proposed to detect anomalous data in public datasets and local hospital children's wellness clinic databases. The proposed framework consists of two main parts. First, the Probabilistic Mapping step consists of k-NN instance acquisition, data distribution calculation, and data point reposition. Truncated Gaussian Distribution (TGD) was used for controlling the boundary of the mapped points. Second, the Outlier Detection step consists of outlier score calculation and outlier selection. Experimental results show that the proposed algorithm outperformed the existing algorithms with real-world benchmark datasets and a Children's Wellness Clinic dataset (CWD). Outlier detection accuracy obtained from the proposed algorithm based on Wellness, Stamps, Arrhythmia, Pima, and Parkinson datasets was 93%, 94%, 80%, 75%, and 72%, respectively.

Keywords: Outlier detection, k-NN, Truncated Gaussian Distribution, Probabilistic Mapped, Mean shift

DOI: 10.37936/ecti-cit.2021152.244971

Article history: received February 17, 2021; revised July 30, 2021; accepted August 5, 2021; available online August 12, 2021

1. INTRODUCTION

Outlier detection was studied in order to enhance the classification ratio of the future data points. In general, an outlier detection problem refers to the distinction between faraway data points and common data points. However, if two classes are located in nearby areas and some parts of the first-class occur in the boundary of the second class, this is called an overlapping problem. In some cases, data points of the overlapping classes are placed in the same positions, but their outliers are marked in the area of the other class. Moreover, the Parkinson dataset contains outliers of about 75%, so it is much more difficult to identify the outlier positions in this dataset. Fig. 1 displays an examples of different characteristic outlier problems. In the example, there are two classes: an outlier and a normal class. The common outlier problem is illustrated in Fig. 1(a). Fig. 1(b) shows the normal class is larger than the outlier class but some parts of the outlier class occur at the boundary of the normal class. In contrast, in Fig. 1(c), the size of the outlier is larger than the normal class and they overlap. Note that in this study the relative sizes of different classes is not the main concern and has no effect on the accuracy of detection.

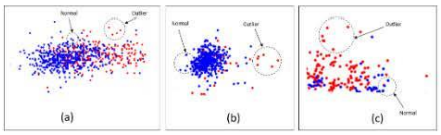


Fig.1: *An example of outlier data (a) common outlier problem (b) normal class is larger than the outlier class (c) the size of the outlier is larger than the normal class.*

In Farag et al. [1][2][3], a novel data clustering algorithm based on gravity centre methodology was introduced. The strength of the gravity centre algorithm is that it does not need to specify parameters. But the disadvantage is that it cannot perform well

[1,2]The authors are with Faculty of Informatics Burapha University, Thailand., E-mail: 61910138@go.buu.ac.th and rasmequa@go.buu.ac.th

[2]Corresponding author: rasmequa@go.buu.ac.th

This work is funded by the faculty of Informatics, Burapha University, fiscal year 2019.

with a very small volume of data. The results of data clustering using Gravity Centre are compared with 3 other methods: K-means, K-medians, and K-medoids. The results indicate that Gravity Centre outperforms those techniques with the datasets used which were NNDSS's family, Health-infectious-disease-2001–2014, Unplanned Hospital Visits – Hospital, Diabetes, and Medicare National DMEPOS HCPCS.

Xiaokang et al. [4], proposed a density-weighted fuzzy outlier clustering approach for class imbalanced learning as a method for clustering fuzzy outlier clusters. This method considers the relationship of new ambiguous neighborhoods with local density data. When weighing the sample in the clustering process, it is mixed with the fuzzy outlier grouping method. In this way, the most representative samples are selected, while the anomalous samples are eliminated. The accuracy of this method shows a superior performance of 92% compared to other cluster sampling models. This indicates that the density-weighted fuzzy outlier grouping method can be used with imbalanced problems in real life. Blood-transfusion, Parkinson, Sick_numeric2, WDBC, and Wine datasets were used in this paper.

In Jiang et al. [5], a local-gravitation-based method for the detection of outliers and boundary points is presented. In it, each data point is viewed as an object with both mass and local resulting force (LRF) generated by its neighbor. When the number of neighbors increases, the LRF of outliers, boundary points, and internal scores change at different rates. The LRF rate of change of a lower density point has a higher scores. That is, the rate of change of the outlier is higher than that of the boundary and internal points. In other words, the highest-ranking score can be identified as an outlier. Likewise, the higher the rate of change of a point, the higher the LRF, and the greater the chance of it being a boundary point. The main advantage of the method is that it is independent of K-value selection. This will result in improving detection efficiency. Heart disease, Lymphography, Ionosphere, Breast cancer Wisconsin, Blood transfusion service center, and SPECTF heart datasets were used.

Aditya and Fitra [6] presented a method for Outlier Detection with a Supervised Learning Method. In their work, several popular classification methods, K-Nearest Neighbor, Centroid Classifier, and Naive Bayes, were compared as tools to handle outlier detection tasks. The results show that those methods achieved 81% of average sensitivity. They report that this is a reasonable starting value for future research to further modify the said methods to improve their performance. Elhossiny et al. [7] proposed using an enhanced K-Means++ to handle missing data and outliers. They worked on a diabetic classification system. They are looking for features that are related to classifying people into two groups: a diabetes group and a control group. The experimental result using RMSE is 17%. Vowels, Thyroid, Vertebral, Wine, Satellite, Breast cancer, and Ionosphere datasets were used.

Diego et al. [8] proposed the DBSCAN technique for large datasets. DBSCAN is a classic clustering method for identifying heterogeneous and isolated clusters with noise. There are a number of articles addressing DBSCAN scalability issues. Despite the scalability issues, the DBSCAN algorithm offers a reduction in execution time due to the reduction in the number of data formats. They also proposed a new heuristic called I-DBSCAN that can be modified and produced good results for the entire data set without any additional parameters. Abalone (Scale), Mushrooms, Pendigits, Letter, Cadata, Shuttle, Sensorless (Scale), SensIT (acoustic), SensIT (seismic), Skin-Nonskin and Poker datasets were used.

Paweł et al. [9] proposed the Fuzzy C-Means based Isolation Forest for outlier detection. In their study, they examined the feasibility of the proposed technique and analyzed it in detail for the different characteristics of data. For example, they considered database size, number of record attributes, and data type. FCM allows membership grade of the generated Isolation Forest nodes to the cluster based on these nodes to be generated. Therefore, this can lead to a fault score for a given element that may belong to a group of similar elements. To overcome this issue, a separate set of forest enrichment methods based on Fuzzy C-Means is proposed. The experimental results performed using 27 datasets indicated that FCM can play a key role in improving forest isolation approaches and increase the value of specific measures on the effectiveness of anomaly detection methods. Annthyroid, Arrhythmia, Breastw, Cardio, Cover, Glass, Ionosphere, Letter, Lympho, Mammography, Mnist, Musk, Optdigits, Pendigits, Pima, Satellite, Satimage-2, Shuttle, Speech, Thyroid, Vertebral, Vowels, Wine, Nad, and unsw0 datasets were used.

Paweł et al. [10] proposed a K-Means-based isolation forest to detect outliers. The K-Means-based Isolation Forest approach allows the creation of search maps. By employing many branches, as opposed to only two as considered in the traditional method, the k-mean grouping is used to estimate the number of divisions on each decision tree node. The proposed method is effective for information coming from different application areas including inter-model transport and spatial data. The advantage of this method is that information can be entered in the process of creating a decision tree. Moreover, it returns a more interesting anomaly score. Artificial sets, NYC Taxi, NYC Taxi (geographical positions), Ship transport, Ship transport, Train transport, Train transport, and Train transport datasets were used.

Patel and Kushwaha [11] reported in "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model" that classification using GMM will enable discovering complex patterns and grouping them into cohesive, homogeneous components that are close representatives of real patterns within the dataset. In other words, it enables high cohesion among members of the same classes. That is, the dependency between the members is high. This can imply that it provides low coupling among different classes or the dependency between the classes is low.

In Yang et. al. [12], a mean-shift outlier detection and filtering technique is proposed to remove the bias caused from a large number of outliers before clustering without the need to know the number of outliers in advance. This method can also be applied with both numeric and string data. The experimental result of this method for the filtering task outperforms five other existing outlier removal methods: LOF, ODIN, NC, IFOREST and ABOD. This method also outperforms a number of the existing outlier detection methods: LOF, NC, KNN, ODIN, MCD, IFOREST, OCSVM, PCAD, and ABOD. This paper used Generated data, KDD-Cup99, Stamps, PageBlocks, Pima, Arrhythmia, and Parkinson datasets in the experiments.

In this work, the Probabilistic Mapped Mean-Shift Algorithm is proposed to detect outliers in public and children's wellness datasets. This method is based on its strong point that there is no need to know the number of outliers in advance. The difference from the classic mean-shift technique is the dispersion of the target position of the mapping depending on the joint probability density function of the mapping function. Unlike the mapping function proposed in [12], our proposed function will map the outliers to the proper point inside the boundary of each class. The mapped position is calculated using a Truncated Gaussian kernel. In other words, each outlier point has been mapped to a new position based on its probability value during the mapping process.

2. PROBLEM ANALYSIS

The framework proposed in this study is based on the conditions illustrated in Fig 1 (b) and (c). Given binary classes, outlier and normal points are occasionary placed in the same positions. We assume that the outlier class may overlap the normal class. There is no constraint imposed on the number of elements between the outlier class (n_o) and the normal class (n_n). This means that both $n_o > n_n$ and $n_n > n_o$ can be cases in this study. The only condition is the range of the output area of the mapping function. The mapping function proposed in [12] tried to map all outlier points into the mean or median point of each cluster as shown in Fig 2(a). In other words, this function will place all outlier points at the center of each cluster. This causes miss-classification in the testing step. Hence, the outlier detection framework proposed in this study focuses on the following issues:

1. How to increase the distribution of the range from the mapping function?
2. How to increase the coverage area of the mapped points obtained in issue 1, as shown in Fig. 2(b)?

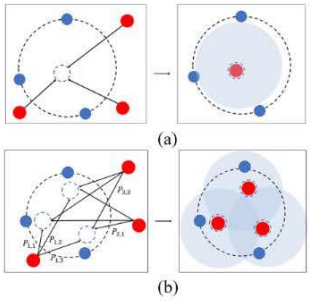


Fig.2: *Mapping functions: (a) Mean Shift Algorithm, (b) Probabilistic Mapped Mean-Shift Algorithm.*

In general, the mean shift algorithm will translate all outlier data points of each cluster to new positions as shown in Fig. 2(a). But in some scenarios, more distribution and more coverage are needed as shown in Fig. 2(b).

3. BACKGROUND

3.1 Mean-Shift Algorithm

The Mean Shift algorithm is a non-parametric clustering algorithm for finding high-density areas (a density function) of the input feature space. It is sometimes called the mode-seeking algorithm. The Mean Shift technique does not need to know the number of clusters in advance. In addition, it is not limited by the shape of the cluster. The algorithm is an iterative procedure. It will repeat until it reaches the optimal area having maxima mode. This algorithm consists of two main steps as follows:

step 1: Compute the mean shift vector: $m(X^t)$
step 2: Translate the input X^t to a new position: $X^t_{map} = X^t + m(X^t)$

3.2 Probabilistic Mapped Model

A probabilistic decision model is a decision method for predicting future events based on probability theory. In contrast, a deterministic decision determines an exact circumstance. The probabilistic mapped model is a mapping function for predicting a new position (Range) of input data (Domain) based on the

its properties. Generally, probabilistic mapped methods will work better than deterministic models. This is because the probabilistic method uses all properties or information contained in the input data. The combination of the probabilistic model with the mean shift method, which is an iterative mapping procedure, help improve the mapping process. That is, a better coverage area for each individual cluster is reached. Hence, outlier prediction accuracy will be higher. This is because the mapping of input data points from the current position is assigned to a new proper area within its class.

4. THE PROPOSED METHOD

In this work, the anomalous data, or outliers, in the large public benchmarks and Children's Wellness Clinic dataset were detected using our Probabilistic Mapped Mean-Shift Algorithm. The proposed method is based on the Mean Shift technique proposed by Yang et. al. [12] as mentioned in Section 1. The proposed framework is divided into two main sections. Section I: Probabilistic Mapping, consists of k-NN instance acquisition, data distribution calculation, and data point reposition. Section II: Outlier Detection, consists of outlier score calculation and outlier selection. A brief description of the proposed method is shown in Fig. 3.

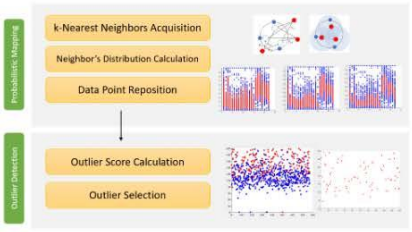


Fig.3: *The proposed framework.*

4.1 Probabilistic Mapping

In this section, mapping the input instance into a new location depends on its probalistic value is described. This procedure consists of three steps. In Step 1, the finding of k-nearest neighbors of each instance is carried out. In Step 2, the calculation of local Gaussian distribution is performed. In Step 3, the repositioning of the instance is achieved.

4.1.1 k-NN Acquisition

To calculate the local Gaussian distribution around the instance, a set of the nearest points is acquired by the k-nearest neighbor algorithm. The steps of k-NN can be described as follows:

step 1: Set value for $k, k \epsilon I^+$
step 2: For each point in the data do the following
→ **2.1** Calculate the distance between data point and each row of the dataset.
→ **2.2** Sort the data in ascending order based on the distance value.
→ **2.3** Select top k data point for the nearest group.
step 3: Repeat these steps until reaching the end of the dataset.

4.1.2 Neighbour's Distribution Calculation

After the data has been separated into groups by k-NN, each group of data is used to calculate local data distribution. To compute local probabilistic distribution, the Truncated Gaussian Distribution (TGD) model was applied. TGD can be described with Eq. (1).

$$(x; \mu, \sigma, a, b) = \frac{1}{\sigma} \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \quad (1)$$

Let X be a normal distribution with mean μ and variance σ^2, within the interval (a,b). Then X conditional on $a \leq x \leq b$ has a Truncated Gaussian Distribution.

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\xi^2\right) \quad (2)$$

ϕ is the probability density function of the standard Gaussian Distribution.

$$\Phi(x) = \frac{1}{2}(1 + \text{erf}(x/\sqrt{2})) \quad (3)$$

Φ is its cumulative distribution function

The steps of TGD mapping 4 can be described as follows:

step 1: Calculate the lower boundary a and upper boundary b of probability in each k-NN set.
step 2: Calculate the Mean μ and standard deviation σ of each k-NN set.
step 3: Calculate distribution probability using Eq. (2).
step 4: Repeat until all k-NN sets have been processed as shown in Fig. 4.

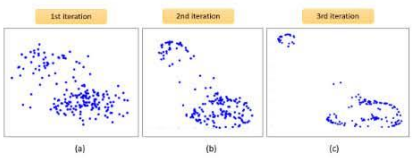


Fig.4: *Mapped data points using Probabilistic Mapped Mean-Shift Algorithms.*

4.1.3 Repositioning Data Points

To shift the data points to their new positions as shown in Fig. 5 for Synthetic data and Fig. 6 for Parkingson data, the highest probability value is selected to assign each new position. In Truncated Gaussian Repositioning there are three main steps:

step 1: Select the initial point from the k-NN set.
step 2: Repeat until converged:
→ **2.1** For each point, find weights encoding the probability of membership in each cluster
→ **2.2** For each cluster, update its location, normalization, and shape based on all data points, making use of the weights.
step 3: Select the highest probability weight.

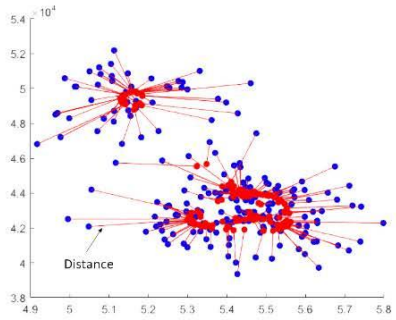


Fig.5: *Visualization of the Mapped points of each cluster in synthetic data. Blue points are original points and Red points are mapped points.*

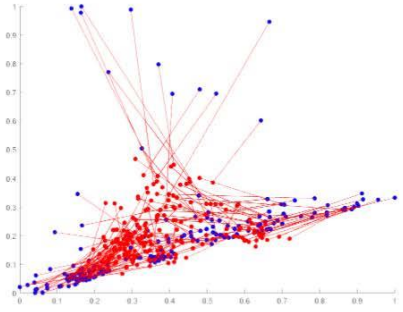


Fig.6: *Visualization of the Mapped points of each cluster in benchmark Parkinson data. Blue points are original points and Red points are mapped points.*

In Fig. 7, 3D space mapped points of Parkinson data are illustrated and the results show that the data was mapped into the new areas with the highest probability. This will increase the cohesion and uniformity between class members for each class as shown in Fig. 8.

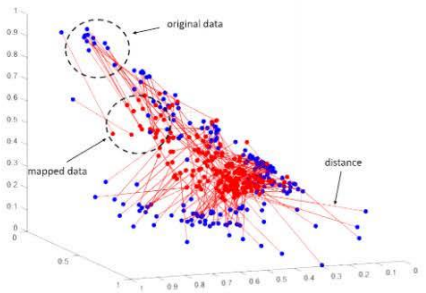


Fig.7: *Visualization in 3D space of the Mapped points of each cluster in benchmark Parkinson data. Blue points are original points and Red points are mapped points.*

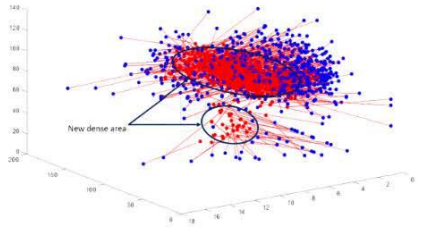


Fig.8: *Resulting Denser Area after Mapping.*

The result shows that the normal data (Blue vertices) was mapped into the denser area (Red vertices). The distribution of each class was normality tested using Shapiro-Wilk test [13]. The result shows the mapped data gained higher normality rates as shown in Eq. (4).

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

Let $x_{(i)}$ be the i^{th} order statistic, such as, the i^{th}-smallest number in the sample. The coefficients a_i are given by Eq.(5).

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{C} \tag{5}$$

Where C is the normalized vector given in Eq.(6).

$$C = \sqrt{(m^T V^{-1} V^{-1} m)}, m \in (m_1, \ldots, m_n)^T \quad (6)$$

4.2 Outlier Detection

In this section, to detect outliers among normal data, two steps are used. In Step 1, the outlier score is calculated. In Step 2, an outlier selection is performed.

4.2.1 Outlier Score Calculation

By using the local distribution of its k-NN, the current object was forcibly repositioned to the denser mean probability area. The length of the object's movement was computed as a piece of outlier evidence. To calculate the outlier score, instead of depending on only clustering or classification techniques, the distance between mapped instances and original data was used. The farther an instance was moved, the stronger the outlier score obtained. The distance between the objects can be calculated with Eq. (7).

$$d(X_i, X_{map(i)}) = \sum_{i=1}^{dn} |X_i - X_{map(i)}| \quad (7)$$

Let $d(X_i, X_{map(i)})$ be the distance between the original data and the mapped instance. X_i is data from the original set and $X_{map(i)}$is data from the mapped instance.

The mapped data contains more information from the data. In the case of a large number of outliers among the normal class, for more robustness, the variant of the distance set is used as an extended reference set for outlier score calculation. Thus, the outlier score can be computed by Eq.(8).

$$S = \sum_{i=1}^{n} |X_i - X_{map(i)}| + \frac{\sum (X_i - \mu)^2}{n-1} \quad (8)$$

Let S be an outlier score. X_i is data from the original set and $X_{map(i)}$ is data from the mapped instance. μ is the mean of the k-NN sets. The score is used in the outlier selection step. The steps for score calculation are:

step 1: Find the k-nearest neighbor for each instance for local variant calculation.
step 2: Compute the outlier score using Eq. (8)
step 3: Repeat until reaching the end of the dataset.

The distance between mapped points and original data points as depicted in Fig. 9 was used in outlier score calculation. Then, those scores were used in the outlier selection step.

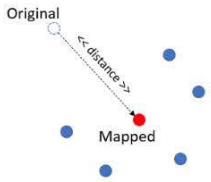


Fig.9: *Example of distance acquisition between mapped data and original data.*

The scores from benchmark datasets, Stamp (Fig. 10 (a)), Pima (Fig.10 (b)), Arrh (Fig.10 (c)), and Parkinson (Fig.10 (d)) are illustrated in Fig. 10.

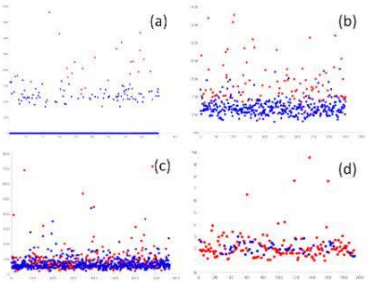


Fig.10: *Visualization of outlier score calculation.*

4.2.2 Outlier Selection

To select outliers from the entire dataset, Yang et. al. [12] proposed a procedure to designate the Top-N outlier score. The number of chosen points can be determined using Eq. (9).

$$Top_n = \left\lfloor \frac{1}{2}(R_{out} N) \right\rfloor, Top_n \in I^+ \quad (9)$$

Top_n is the number of Top-N outlier scores. R_{out} is an outlier ratio or percentage of the datasets. N is the number of entire instances. For example, the Parkinson dataset contains 75.4% outliers. The Top-N of Parkinson dataset can be calculated by $\frac{1}{2}(0.75)(195) \approx 73$ points as shown in Table 1.

Table 1: *Top-N Selection Results.*

Dataset	Instances	Top-N
CWC	2533	50
Stamps	340	19
Arrhythmia	450	36
Pima	768	138
Parkinson	195	73

In Fig.11, Parkinson outlier Scores are shown. The red dots and blue dots are likely outliers and normal data respectively. In this step, the calculated outlier scores without sorting are shown in Fig. 11 (a). In order to select outlier candidates with the Top-N approach, the outlier scores must be sorted in descending order as shown in Fig. 11 (b). The outlier selection steps are:

step 1: Calculate Top-N for selecting outlier candidates using Eq. (9)

step 2: Sort the data points with outlier scores in descending order as shown in Fig.11 (b)

step 3: Select the first Top-N data points (from **step 1**) for outlier candidate selection.

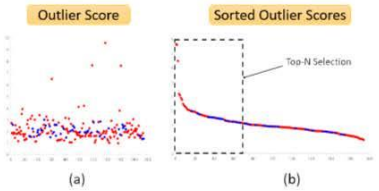


Fig.11: *Visualization of Parkinson Outlier Scores.*

The procedure of the Probabilistic Mapped Mean-Shift Outlier Detection (PMMS) process is summarised in Algorithm 1.

Algorithm 1 Probabilistic-Mapping Mean-Shift $\{PMMS(X,k,i)\}$
Input: Dataset X, Nearest Neighbor k, Iteration i
Output: Mapped Dataset X_{map}, Outlier Score S

REPEAT i **TIMES:**
  FOR $X_i \in X$:
    $K_{nn} \leftarrow kNN(X_i), K_{nn} \in X$ Find k-nearest neighbors of X_i
    COMPUTE the probabilistic distribution of K_{nn}
      $a_i \leftarrow$ **COMPUTE** lower boundary of K_{nn}
      $b_i \leftarrow$ **COMPUTE** upper boundary of K_{nn}
      $\sigma_i \leftarrow$ **COMPUTE** Standard Deviation of K_{nn}
      $\mu_i \leftarrow$ **COMPUTE** Mean of K_{nn}
    COMPUTE Truncated Gaussian Distribution
    FOR $k_j \in K_{nn}$:

$$P_i(k_j;\mu_i,\sigma_i,a_i,b_i) = \frac{1}{\sigma_i} \frac{\phi\left(\frac{k_j-\mu_i}{\sigma_i}\right)}{\Phi\left(\frac{b_i-\mu_i}{\sigma_i}\right)\Phi\left(\frac{a_i-\mu_i}{\sigma_i}\right)}$$

    $X_{map} \leftarrow P_i$ **REPOSITION** X_i **WITH** P_i
COMPUTE Outlier Score
  FOR $X_i \in X$:
    $K_{nn} \leftarrow kNN(X_i), K_{nn} \in X$ Find k-nearest neighbors of X_i
    $S_i \leftarrow \sum_{o=1}^{n} |X_i - X_{map(o)}| + \frac{\sum(X_i-\mu)^2}{n-1}$, $X_{map(o)} \in X_{map}$, $X_i \in X$

5. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the dataset used in this experiment and a performance evaluation of the proposed methods are described.

5.1 Dataset

The proposed approach was tested on four public benchmark datasets and a real-world children's wellness clinic dataset.

5.1.1 Public Datasets

The proposed method was tested against the public datasets Stamp, Arrhythmia, Pima, and Parkinson as performed in Yang et. al. [12]. Those public datasets (Stamps, Arrh, Pima, and Parkinson) have outlier ratios of 9.1%, 15%, 34.9%, and 75.4%, respectively as shown in Table 2. The detection accuracy, as shown in Table 3, obtained from the existing methods is 89%, 73%, 74%, and 68%, respectively. These accuracy rates left room for improvement.

5.1.2 Children's Wellness Clinic dataset (CWC)

The real-world dataset was provided from Children's Wellness Clinic of Burapha University Hospital. It consists of 2533 instances with 9 dimensions. The outliers were labelled as 4% of entire dataset.

Table 2: *Benchmark and Real World dataset description.*

Dataset	Instances	Dimension	Outlier(%)
CWC	2533	9	4
Stamps	340	9	9.10
Arrhythmia	450	259	15.8
Pima	768	8	34.9
Parkinson	195	22	75.4

5.2 Experimental Results

The experimental results revealed in Table 3 indicate that these results address the first research issue concerning how the large dataset effects detection of outliers. For the second issue, the experimental results show that adding additional parameters helps improve the detection performance. That is, the proposed method with adjusted parameters outperforms the existing methods. In the performance evaluation, the accuracy is obtained with the confusion matrix measurement in Eq. (10).

$$Acc = \frac{TP}{TP + FP} \tag{10}$$

Let TP be the number of True positive cases (number of outliers were selected in Top-N method). FP is the number of False Negative cases (number of normal points were selected in Top-N method). The result shows the proposed method obtained 0.94, 0.80, 0.75, 0.72 accuracies in benchmark datasets, consists of Stamps, Arrhythmia, Pima, and Parkinson, respectively as shown in Table 3 and Fig. 12.

Table 3: *BResults of the proposed method compared with the existing methods for detecting anomalous data.*

Dataset ((Outliers)	CWC. (4%)	Stamps (9.1%)	Arrh (15%)	Pima (34.9%)	Parkinson (75.4%)	Average
PMMS (proposed)	**0.93**	**0.94**	**0.80**	**0.75**	**0.72**	**0.80**
DOD+	0.91	0.89	0.73	0.71	0.63	0.74
MOD+	0.89	0.91	0.72	0.73	0.66	0.76
DOD	0.92	0.81	0.73	0.68	0.68	0.73
MOD	0.92	0.78	0.74	0.65	0.60	0.69
LOF	-	0.53	0.73	0.60	0.56	0.61
ODIN	-	0.58	0.70	0.56	0.45	0.57
NC	-	0.50	0.60	0.52	0.46	0.52
KNN	-	0.89	0.74	0.72	0.54	0.72
ABOD	-	0.81	0.72	0.70	0.64	0.72

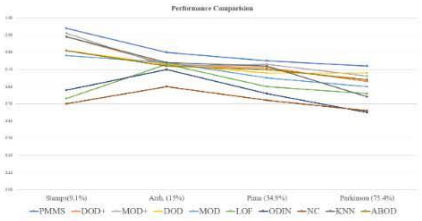


Fig.12: *Performance Comparation.*

Likewise, the proposed method also outperforms the existing methods for the real world dataset Children's Wellness Clinic with an accuracy rate of 0.93. However, the proposed method needs to be improved further to reach higher accuracy in large datasets and datasets containing a large amount of outlier data.

5.3 Discussion

Although the proposed method outperforms the others (see Table 3), the accuracy is inverse to the outlier percentage as shown in Fig. 13. In the other words, the accuracy of the proposed approach tends to be lower in cases with a larger percentage of outliers. For the Parkinson dataset, which contains 75% outliers, the proposed approach obtains the lowest accuracy rate. To eliminate this weakness, an improved probability model and better outlier selection methods need to be found.

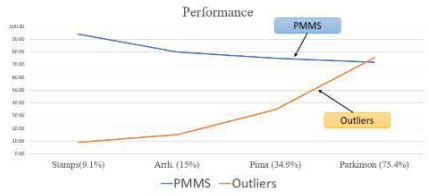


Fig.13: *Relationship between proposed method and percentage of outliers.*

In addition, not only the datasets having high percentages of outliers decrease the accuracy, but the datasets having imbalanced class members do also. The latter case is shown in Fig. 14. Thus, in the future work, the detection of outliers for imbalanced classes will be pursued.

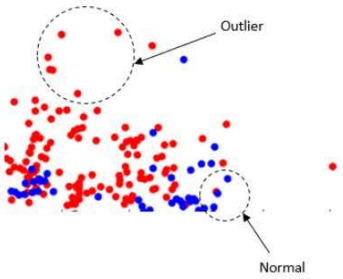


Fig.14: *Visualization of fully overlapped problem.*

6. CONCLUSION

To enhance machine-learning or data-mining algorithms such as classification and clustering, outlier data must be eliminated. In this research, the Probabilistic Mapped Mean-Shift Algorithm is proposed to detect and filter outlier data in public benchmark and local hospital children's wellness clinic datasets. The experiment results indicated that the proposed method can address the first research issue by increasing the distribution of the range. The results also demonstrated that the proposed method can address the second issue by increasing the coverage area of the mapped points obtained in issue 1, so that each group of mapped points has its own normality. The normality of shifted instances was evaluated using the Shapiro-Wilk test. The test result demonstrated that the normality rate is increased. However, the distribution and the coverage still need to be improved further. In addition a wider variety of datasets should be used to test our method.

The proposed approach consists of two main phases: Phase I, Probabilistic Mapping, including k-NN instance acquisition, data distribution calculation, and data point reposition, and Phase II: Outlier Detection, including outlier score calculation, and outlier selection. The proposed approach was tested on benchmark datasets including Stamps, Arrhythmia, Pima, and Parkinson datasets, and also on a local hospital Children's Wellness Clinic (CWC) dataset. Experimental results indicated that the proposed algorithm achieves higher accuracy than the existing algorithms. The accuracy achieved by the proposed method is 94%, 80%, 75%, 72%, and 93% respectively. In future work, datasets with fully overlapped and imbalanced classes, as shown in Fig. 14, will be taken into consideration.

References

[1] A. Boukerche, L. Zheng and O. Alfandi, "Outlier Detection: Methods, Models, and Classification," *ACM Computing Surveys (CSUR)*, Vol.53, pp.1-37, 2020.

[2] J. Huang, Q. Zhu, L. Yang, D. D. Cheng and Q. Wu, "A novel outlier cluster detection algorithm without top-n parameter," *Knowledge-Based Systems*, Vol.121, pp.32-40, 2017.

[3] F. H. Kuwil, Ü. Atila, R. Abu-Issa and F. Murtagh,, "A novel data clustering algorithm based on gravity center methodology," *Expert Systems with Applications*, Vol. 156, 2020.

[4] X. Wang, H. Wang and Y. Wang, "A density weighted fuzzy outlier clustering approach for class imbalanced learning," *Springer Neural Computing and Applications*, 2020.

[5] J. Xie, Z. Xiong, Q. Dai, X. Wang and Y. Zhang, "A local-gravitation-based method for the detection of outliers and boundary points," *Knowledge-Based Systems*, Vol.192, 2020.

[6] A. H. Bawono and F. A. Bachtiar, "Outlier Detection with Supervised Learning Method," *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, pp.306-309, 2019.

[7] E. Ibrahim, M. A. Shouman, H. Torkey and A. El-Sayed, "Handling missing and outliers values by enhanced algorithms for an accurate diabetic classification system," *Springer Multimedia and Applications*, Vol.80, pp.20125-20147, 2021.

[8] D. Luchi, A. L. Rodrigues, F. M. Varejão, "Sampling approaches for applying DBSCAN to large datasets," *Pattern Recognition Letters*, Vol. 117, pp.90-96, 2019.

[9] P. Karczmarek, A. Kiersztyn, W. Pedrycz and E. Al, "K-Means-based isolation forest," *Knowledge-Based Systems*, Vol. 195, 2020.

[10] P. Karczmarek, A. Kiersztyn, W. Pedrycz and D. Czerwiński,, "Fuzzy C-Means-based Isolation Forest," *Elsevier Applied Soft Computing*, Vol.106, 2021.

[11] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," *Procedia Computer Science*, Vol.171, pp.158-167, 2020.

[12] J. Yang, S. Rahardja and P. Fränti, "Mean-shift outlier detection and filtering," *Pattern Recognition*, Vol.115, 2021.

[13] E. González-Estrada and W. Cosmes, "Shapiro–Wilk test for skew normal distributions based on data transformations," *Journal of Statistical Computation and Simulation*, Vol.89, Issue 17, pp.3258-3272, 2019.

Siriwan Phongsasiri received the B.Sc. degree in Computer Science from Burapha University, Chonburi, Thailand, in 2017. Currently, she is a M.Sc. student in Informatics, at Faculty of Informatics, Burapha University, Chonburi, Thailand.

Suwanna Rasmequan received the B.B.A. degree in Finance and Banking and the M.Sc. degree in Computer Information System from Assumption University, Bangkok, Thailand, in1992 and 1994 respectively. She received Ph.D. degrees in Computer Science in 2002 from University of Warwick, Coventry, United Kingdom. She worked in the Business Sector from 1984 until 1997 in two major businesses namely Packaging Industry and Satellite Communication Provider. Her work responsibilities in those businesses was starting with the beginning post of Executive Secretary and ending with the post of Section Manager. She was a Lecturer at the Department of Computer Science, Burapha University, Chonburi, Thailand during 1997 – 2006. She has been worked as an Assistant Professor from year 2006 up to present at the Faculty of Informatics, Burapha University. Her major research interests include Empirical Modelling, Decision Support Systems, Machine Learning Applications.

บรรณานุกรม

Bansal, S. and N. Baliyan (2019). "A Study of Recent Recommender System Techniques." International Journal of Knowledge and Systems Science 10(2): 13-41.

Boukerche, A., et al. (2020). "Outlier Detection." ACM Computing Surveys 53(3): 1-37.

Cheng, T. (2017). An Improved DBSCAN Clustering Algorithm for Multi-density Datasets. Proceedings of the 2nd International Conference on Intelligent Information Processing - IIP'17: 1-5.

Cheng, Z., et al. (2019). Outlier detection using isolation forest and local outlier factor. Proceedings of the Conference on Research in Adaptive and Convergent Systems: 161-168.

Huang, J., et al. (2017). "A novel outlier cluster detection algorithm without top-n parameter." Knowledge-Based Systems 121: 32-40.

Karczmarek, P., et al. (2020). "K-Means-based isolation forest." Knowledge-Based Systems 195.

Karczmarek, P., et al. (2021). "Fuzzy C-Means-based Isolation Forest." Applied Soft Computing 106.

Kbaier, M. E. B. H., et al. (2017). A Personalized Hybrid Tourism Recommender System. 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA): 244-250.

Kouadria, A., et al. (2019). "A Multi-criteria Collaborative Filtering Recommender System Using Learning-to-Rank and Rank Aggregation." Arabian Journal for Science and Engineering 45(4): 2835-2845.

Kuwil, F. H., et al. (2020). "A novel data clustering algorithm based on gravity center methodology." Expert Systems with Applications 156.

Lafta, R., et al. (2015). An Intelligent Recommender System Based on Short-Term Risk Prediction for Heart Disease Patients. 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT): 102-105.

Luchi, D., et al. (2019). "Sampling approaches for applying DBSCAN to large datasets." Pattern Recognition Letters 117: 90-96.

Touati, R., et al. (2020). "Anomaly Feature Learning for Unsupervised Change Detection in Heterogeneous Images: A Deep Sparse Residual Model." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13: 588-600.

Xie, J., et al. (2020). "A local-gravitation-based method for the detection of outliers and boundary points." Knowledge-Based Systems 192.

Xu, D., et al. (2017). An Improved Data Anomaly Detection Method Based on Isolation Forest. 2017 10th International Symposium on Computational Intelligence and Design (ISCID): 287-291.

Yang, J., et al. (2021). "Mean-shift outlier detection and filtering." Pattern Recognition 115.