การสร้างแบบจำลองจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรังโดยใช้เทคนิคเหมืองข้อมูลและวิชวลไลเซชั่น

Data Classification Model for Chronic Kidney Disease using data mining Technique and Visualization

ประยุทธ์ศิลป์ ชัยนาม

สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม

E-mail: prayutsilp@yahoo.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบโมเดลสำหรับจำแนกผู้ป่วยโรคไตเรื้อรัง โดยอาศัยเทคโนโลยี การทำเหมืองข้อมูล โดยเลือกใช้ วิธีต้นไม้ตัดสินใจ วิธีต้นไม้ตัดสินใจแบบสุ่ม วิธีความใกล้เคียงกันด้วยค่าเค วิธีซัพพอร์ตเวคเตอร์แมชชีน วิธีนาอีฟเบย์ มาใช้ในการทดสอบกระจำแนกกลุ่มผู้ป่วยโดยใช้ข้อมูลผู้ป่วยโรคไตเรื้อรังของโรงพยาบาลอพอลโล ประเทศอินเดีย โดยเก็บไว้ในฐานข้อมูลเชิงสัมพันธ์ ซึ่งมีประสิทธิภาพที่ดีสามารถรองรับการส่งผ่านข้อมูลได้อย่างรวดเร็วซึ่งตอบโจทย์การใช้งานข้อมูลสารสนเทศในปัจจุบัน และวิเคราะห์เพื่อสรุปผลด้วยระบบวิเคราะห์ข้อมูลเชิงรูปภาพเพื่อให้สามารถเข้าใจข้อมูลได้ง่าย

คำสำคัญ: วิชวลไลเซชั่น เหมืองข้อมูล ต้นไม้ตัดสินใจ ค่า เค ใกล้เคียง ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์

Abstract

This research aims to develop and compare models for identifying chronic kidney disease patients. By using serveral Data mining techniques including K-nearest neighbor, decision tree, Random Forest, support vector machine and Naïve Bayes are used to test the classification of patients by using chronic kidney disease data from Apollo Hospital India By storing it in a relational database. Which has good efficiency, can support fast data transmission which responds to current information usage. And analyze to summarize with the image analysis system to be able to understand the information easily.

Keywords: Visualization, Data Mining-nearest neighbor, decision tree, artificial neural network, support vector machine, Naïve Bayes

1. ที่มาและความสำคัญ

ปัจจุบันข้อมูลมีอยู่เป็นจำนวนมากโดยเฉพาะองค์กรขนาดใหญ่อย่างโรงพยายาลที่มีข้อมูลประวัติของคนไข้จำนวนมหาศาล รวมไปถึงข้อมูลการรักษาของผู้ป่วยที่ต้องมีการบันทึกข้อมูลอย่างละเอียดแม่นยำ นอกจากนี้ยังมีข้อมูลของยาที่ใช้รักษาที่มีหลากหลายชนิดเช่นกันซึ่งจะทำให้เกิดข้อมูลสะสมจำนวนมาก ซึ่งการเก็บข้อมูลเหล่านี้ลงในฐานข้อมูลแบบทั่วไปไม่สามารถจัดการกับข้อมูลเหล่านี้ได้อย่างมีประสิทธิภาพ เนื่องจากใช้เวลาลานานในการเรียกข้อมูลออกมาวิเคราะห์จึงได้เกิดเทคโนโลยีแบบใหม่เพื่อช่วยให้การจัดเก็บข้อมูลทำได้อย่างมีประสิทธิภาพ การวิเคราะห์ข้อมูลขนาดใหญ่ถือเป็นสิ่งสำคัญในการช่วยให้การดูแลรักษาทำได้ถูกต้องและแม่นยำยิ่งขึ้น ข้อมูลสารสนเทศนั่นจำเป็นต้องมีความถูกต้อง รวดเร็วและแม่นยำ ซึ่งการวิเคราะห์ข้อมูลเหล่านี้หากวิเคราะห์ด้วยวิธีที่ไม่ถูกต้องผิดวิธีก็จะทำให้ผลลัพธ์นั่นมีความถูกต้องลดลง ไม่แสดงถึงผลลัพธ์ที่แท้จริง ดังนั่นผู้วิจัยจึง

ต้องทดสอบวิธีการจำแนกผู้ป่วยให้เหมาะสมกับกลุ่มผู้ป่วยโรคไตเรื้อรัง โรคไตเรื้อรังเป็นโรคที่มีอาการแย่ลงอย่างช้าๆ ทำให้ผู้ป่วยหลายคนไม่ทราบว่าเองป่วยเป็นโรคไตเรื้อรัง ดังนั้นการตรวจพบและได้รับการรักษาให้เร็วที่สุดจึงเป็นสิ่งสำคัญ หากได้รับการรักษาตั้งแต่อาการยังไม่อยู่ในระดับที่ร้ายแรง อาจสามารถทำให้โดรคไตเรื้อรังไม่มีอาการที่แย่ลงหรืออาจหายได้

จากปัญหาดังกล่าวนี้ ผู้วิจัยจึงได้นำ "ระบบวิเคราะห์ข้อมูลเชิงรูปภาพ (Data Visualization) มาใช้ ในการศึกษาครั้งนี้ และได้มีการพัฒนากระบวนการสำหรับคัดแยกและจัดการกับข้อมูลผ่านการใช้ภาษา Python ในการทำ ETL ก่อนจะวิเคราะห์ข้อมูลโดยการใช้ Python (Scikit Learn) และใช้กระบวนการจัดเก็บข้อมูลในรูปแบบของ Relational Database จากนั่นจัดทำข้อมูลภาพเชิงวิเคราะห์ (Data Visualization) โดยจะนำมาใช้ในการเปรียบเทียบและวิเคราะห์ข้อมูล เพื่อเปรียบเทียบผลการจำแนกของแต่ละโมเดลเพื่อให้ได้โมเดลที่ดีที่สุดและจัดทำวิชวลแดชบอร์ด (Visual Dashboard) ที่แสดงการรายงานผลเชิงเปรียบเทียบโดยใช้แผนภาพกราฟฟิกเพื่อประกอบการวางแผนและตัดสินใจในการรักษาผู้ป่วยต่อไป

2. วัตถุประสงค์

2.1 เพื่อศึกษาและสร้างแบบจำลองเหมือข้อมูลที่เหมาะสำหรับการจำแนกผู้ป่วยโรคไตเรื้อรัง

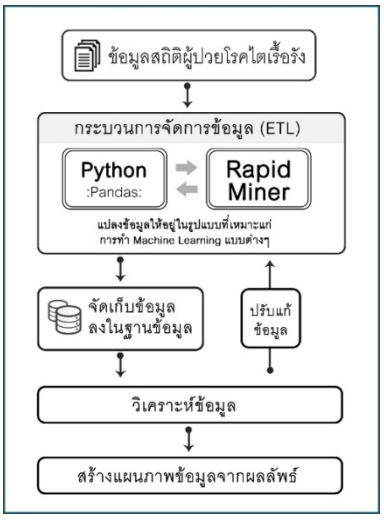
2.2 เพื่อทดสอบและเปรียบเทียบประสิทธิภาพของโมเดลแต่ละประเภท

2.3 ศึกษาการทำวิชวลไลเซชั่นเพื่อให้ง่ายต่อการเข้าใจและตัดสินใจ

3. ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

3.1 การทำเหมืองข้อมูล (Data Mining)

คือกระบวนการที่กระทำต่อข้อมูลจำนวนมากเพื่อจำแนกรูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้นโดยการใช้กระบวนการทาง สถิติ คณิตศาสตร์ และ สมาการ ทำให้เกิดองค์ความรู้ใหม่หรือค้นพบข้อมูลที่น่าสนใจ สามารถนำมาประยุกต์ใช้ให้เกิดประโยชน์ได้ในหลายๆด้าน เช่น ด้านธุรกิจสามารถใช้เทคนิคดาต้าไมนิ่งเพื่อค้นหาหรือคาดการ์ณความต้องการของลูกค้าในอนาคตหรือวิเคราะห์ข้อมูลเพื่อเป็นแนวทางในการออกแบบผลิตภัณฑ์แบบใหม่ให้ตรงตามความต้องการของผู้บริโภค และยังสามารถนำมาใช้วางแผนการใช้ทรัพยากรณ์ขององค์กรเพื่อนำมาประกอบการตัดสินใจของผู้บริหาร



รูปที่ 1 กระบวนการทำเหมืองข้อมูล

3.2 ตัดสินใจแบบต้นไม้ (Decision Tree)

เป็นการเรียนรู้เพื่อแยกประเภทของข้อมูล (Classification) โดยใช้โมเดลแบบต้นไม้ ซึ่งคุณลักษณะ (Attribute) จะเป็นเงื่อนไขข้อกำหนดในการตัดสินใจของต้นไม้ โดยจะมีตัวแปรต้นเป็นตัวแปรที่มีความสำคัญมากที่สุดตัดสินใจในลำดับแรกและเมื่อได้ผลลัพธ์มาก็จะแตกออกมาเป็นกิ่งย่อยเพื่อตัดสินใจด้วยคุณลักษณะที่เป็นเงื่อนไขลำดับถัดมา ต้นไม้ตัดสินใจเป็นอัลกอริธึมที่ถูกนำมาใช้ในหลายๆด้าน ในด้านของธุรกิจสามารถนำมาช่วยในการตัดสินใจเพื่อช่วยในการวางแผนสำหรับโปรเจคเพื่อสร้างแผนงาน และยังสามารถนำมาใช้ในทางการแพทย์เพื่อวินิจฉัยโรคได้อีกด้วย

3.3 Support vector machine (SVM)

เป็น machine learning สำหรับจำแนกข้อมูล โดยใช้สมาการเส้นตรงในการจำแนกข้อมูลที่กระจายตัวและหาเส้นสมการเส้นตรงที่เหมาะสมที่สุดเพื่อหาระนาบแล้วแยกข้อมูลแต่ละกลุ่มออกจากกันและเพื่อหาเส้นตรงที่ดีที่สุดจะมีการหา margin ออกไปทั้ง 2 ฝั่งของเส้นเพื่อหาระยะที่เหมาะสมที่สุดโดยการขยายเส้นขอบจนกว่าจะไปสัมผัสกับข้อมูลโดยที่เส้นขอบที่ขยายออกไปนั่นจะขนานก็เส้นเดิมเพื่อไม่ให้เกิดการ overfitting ของข้อมูล

การจำแนกข้อมูลแบบหลายมิติ จะใช้การเลือกคุณลักษณะที่มีความเหมาะสมที่สุดเรียกว่า โครงสร้างในการคัดเลือก (feature selection) ซึ่งโครงสร้างในการคัดเลือกข้อมูลตัวอย่างที่ใช้ให้ระบบเรียนรู้ จุดมุ่งหมายของSVM คือ แบ่งแยกกลุ่มของเวกเตอร์ในกรณีนี้ด้วยหนึ่งกลุ่มของตัวแปรเป้าหมายที่อยู่ข้างหนึ่งของระนาบ และกรณีของกลุ่มอื่นที่อยู่ทางระนาบต่างกัน ซึ่งเวกเตอร์ที่อยู่ข้างระนาบหลายมิติทั้งหมดเรียกว่า ซัพพอร์ตเวกเตอร์(Support Vectors)

3.4 การเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning)

เป็น machine learning สำหรับคัดแยกปรเภท (Classification) ที่มีความซับซ้อนไม่มากและเป็นที่นิยมในการนำมาวิเคราะห์ข้อมูล โดยอ้างอิงจากทฤษฎีของ เบย์ (Bayes) โดยการใช้หลักการของความน่าจะเป็น (Probability) โดยอาศัยลักษณะ (Feature) หลายๆอย่างมาคำนวนเพื่อหาความน่าจะเป็นและการแจกแจงความน่าจะเป็นตามตามสมมติฐานที่ได้ตั้งไว้เมื่อคำนวนแล้วจะแจกแจงเพื่อนำมาปรับเพิ่มหรือลดลักษณะของข้อมูลโดยรวมเข้ากับข้อมูลเดิมให้เหมาะสมกับตัวโมเดลและให้ได้ค่าความน่าจะเป็นที่ดีที่สุด โดยหลักการของยาอีฟเบย์จะคำนวนความน่าจะเป็นเพื่อจำแนกข้อมูลแบบคาดการ์ณผลลัพธ์ได้และจะวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรร์และนำไปใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ เนื่องจากเป็นอัลกอริทึมที่มีความซับซ้อนไม่มากนักจึงเหมาะกับดาต้าเซ็ตที่มีข้อมูลจำนวนมากและยังเหมาะกับข้อมูลที่มีลักษณะ (Feature) ไม่ขึ้นต่อกันอีกด้วย

3.5 ค่า เค ใกล้เคียง (k nearest neighbor : KNN)

เป็นอัลกอริทึมแบบมีผู้สอน (Supervised) เพื่อใช้สำหรับการจำแนกประเภท (Classification) จัดแบ่งคลาสโดยการวัดระยะห่างของข้อมูลที่ใส่เข้าไปกับข้อมูลที่มีอยู่แล้วด้วยสมการหาระยะทางระหว่างจุดของปีทาโกรัส เพื่อหาจุดข้อมูลใกล้เคียงที่ใกล้ที่สุด โดยจะใช้ค่า K หรือจำนวนข้อมูลในระยะใกล้เคียงที่ต้องการจะใช้ในการตัดสินประเภทของข้อมูล ค่า k จึงเป็นค่าสำคัญที่กำหนดประเภทของข้อมูลได้ ค่า k ที่ต่างกันจะส่งผลกับผลลัพธ์ที่ได้ เนื่องจากค่า k ยิ่งมากจะทำให้ค่าความแปรปรวน (Noise) น้อยลงแต่จะส่งผลให้สูญเสียความแม่นยำไป

3.6 ป่าไม้แบบสุ่ม (Random Forest)

เป็นการเรียนรู้เพื่อแยกประเภทของข้อมูล (Classification) โดยใช้โมเดลแบบต้นไม้ แบบเดียวกับแบบจำลองต้นไม้ตัดสินใจแต่ทำการสุ่มคุณลักษณะหลายๆรูปแบบเพื่อสร้างต้นไม้ตัดสินใจหลายๆต้น ต้นไม้ตัดสินใจแต่ละต้นจะทำการพายกรณ์แบบแยกต่อกันและเลือกผลลัพธ์ของต้นที่มีค่าความแม่นยำมากที่สุด

3.7 การวิเคราะห์ข้อมูลเชิงรูปภาพ (Data Visualization)

การสร้างภาพจากข้อมูลเป็นสิ่งที่จะช่วยให้ผู้คนเข้าใจถึงความสำคัญของข้อมูลโดยการวางไว้ในบริบทของภาพ รูปแบบแนวโน้มและความสัมพันธ์ที่อาจไม่ถูกตรวจพบในข้อมูลที่เป็นสิ่งที่สามารถแสดงได้ง่ายขึ้นด้วยซอฟต์แวร์การสร้างภาพข้อมูล เครื่องมือสร้างภาพข้อมูลสามารถใช้งานได้หลายวิธี การใช้งานที่พบบ่อยที่สุดคือเครื่องมือการรายงานสำหรับธุรกิจ (Bussiness Intelliegnce) ผู้ใช้สามารถตั้งค่าเครื่องมือสร้างภาพข้อมูลเพื่อสร้างแดชบอร์ด(Dash Bord) ได้อัตโนมัติตามความต้องการขององค์กร ผ่านตัวบ่งชี้ประสิทธิภาพหลักและตีความผลลัพธ์ด้วยสายตา

3.8 โรคไตเรื้อรัง (Chronic Kidney Disease)

โรคไตเรื้อรัง คือ สภาวะของไตที่เสื่อมสภาพหรือถูกทำลาย ทำให้ความสารมารถในการจัดการและคัดกรองของเหลวในร่างกายลดลง เช่น การรักษาสมดุลของของเหลวในร่างกาย การควบคุมปริมาณน้ำหรือแร่ธาตุต่างๆในเลือด การกำจัดของเสียจากร่างกายหรือพิษต่างๆออกจากร่างกาย เป็นต้น โดยสาเหตุที่ก่อให้เกิดโรคไตเรื้อรังคือ เบาหวาน ความดันโลหิตสูง และโรคอ้วน รวมถึงสภาวะอื่นๆ เช่น ไตอักเสบ โรคถุงน้ำในไต เป็นต้น

ในระยะแรกของโรคไตเรื้อรังอาจมีอาการหรืออาการแสดงน้อย โรคไตเรื้อรังอาจไม่ปรากฏชัดเจนจนกว่าไตจะบกพร่องอย่างมีนัยสำคัญ การรักษาโรคไตเรื้อรังมุ่งเน้นไปที่การชะลอการลุกลามของความเสียหายของไตโดยการควบคุมสาเหตุพื้นฐาน โรคไตเรื้อรังสามารถพัฒนาไปสู่ภาวะไตวายในระยะสุดท้ายซึ่งเป็นอันตรายถึงชีวิตโดยไม่ต้องมีการกรองเทียม (ล้างไต) หรือการปลูกถ่ายไต

สัญญาณและอาการของโรคไตมักจะไม่ชัดเจนซึ่งหมายความว่าอาการของผู้ป่วยอาจเกิดจากโรคอื่น ๆ เนื่องจากไตของคุณสามารถปรับตัวอย่างมากและสามารถชดเชยการทำงานที่สูญเสียไปอาการและอาจไม่แสดงอาการปรากฏออกมาจนกว่าจะเกิดความเสียหายที่ไม่สามารถกลับคืนสภาพเดิมได้

ระยะของโรค	รายละเอียดของระยะต่าง ๆ	ค่าการทำงานของไต (GFR)
ระยะที่ 1	ไตเริ่มเสื่อม (มีโปรตีนในปัสสาวะ) ค่า GFR ปกติ	90 หรือมากกว่า
ระยะที่ 2	ไตเสื่อม ค่า GFR ลดลงเล็กน้อย	60-89
ระยะที่ 3	ค่า GFR ลดลงปานกลาง	30-59
ระยะที่ 4	ค่า GFR ลดลงมาก	15-29
ระยะที่ 5	ไตวาย	น้อยกว่า 15

รูปที่ 2 ระดับการเสื่อมของไต

ที่มา https://www.bumrungrad.com

3.9 การประเมินโมเดล

ครอสวาลิเดชั่น (Cross Validation) คือวิธีการในการคาดการค่าความผิดพลาดของโมเดลหรือ วิธีการที่เรานำเสนอ โดยพื้นฐานของวิธีการครอสวาลิเดชั่นคือ การสุ่มตัวอย่าง (Resampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆและนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำครอสวาลิเดชั่นมักถูกใช้เป็นตัวเลือกในการกำหนดโมเดล เช่น สถาปัตยกรรมเครือข่ายการสื่อสาร (Network architecture), โมเดลในการคัดแยกประเภท(Classification model) เช่นในการทำ Classify ข้อมูลโดยใช้เทคนิคของ Data mining เช่น Neural Network หรือ Decision Tree นั้นจะต้องมีการแบ่งข้อมูลออกเป็นชุดสอนและชุดทดสอบ แต่ในบางครั้งอาจเกิดปัญหาจากการเลือกข้อมูลที่ดีและ ง่ายมาเป็นข้อมูลชุดทดสอบทำให้ผลการ Classify นั้นดีเกินจริง ดังนั้นจะมีการคิดวิธี k-fold cross validation ขึ้นมาแก้ปัญหาและใช้การวัดความแม่นยำและประสิทธิภาพด้วยการทำ Confuion matrix คู่กับ Classification Report และใช้เทคนิคการวิเคราะห์ข้อมูลด้วยการทำวิชวลไลเซชั่นเพื่อให้เข้าใจถึงความหมายและความสำคัญของคุณลักษณะของกลุ่มผู้ป่วย

4. วิธีดำเนินการวิจัย

4.1 การรวบรวมข้อมูล

ผู้วิจัยได้นำข้อมูลผู้ป่วยโรคไตเรื้อรังจากโรงพยาบาลอพอลโล ประเทศอินเดีย ของ Dr.P.Soundarapandian.M.D.,D.M (Senior Consultant Nephrologist), Apollo Hospitals, Managiri, Madurai Main Road, Karaikudi, Tamilnadu, India. มาใช้ในการวิจัยครั้งนี้

4.2 การเตรียมข้อมูล

ขั้นตอนการเตรียมข้อมูลให้เหมาะสมเพื่อให้สามารถนำไปใช้ในการทำเหมืองข้อมูล มีขั้นตอนดังต่อไปนี้

การคัดกรองข้อมูล (Data Cleaning) เนื่องจากข้อมูลผู้ป่วยโรคไตเรื้อรังบางส่วนมีข้อมูลที่ขาดหาย ข้อมูลที่ซ้ำกัน ข้อมูลบางส่วนมีความผิดปกติ ผู้วิจัยจึงต้องนำข้อมูลเหล่านั่นออกและทำข้อมูลให้ถูกต้อง เพื่อให้การสร้างโมเดลการจำแนกผู้ป่วยมีความถูกต้องและแม่นยำ

การแปลงข้อมูล (Data Transformation) ข้อมูลบางส่วนจำเป็นต้องมีการเปลี่ยนลักษณะของข้อมูลเพื่อให้เหมาะสมและสามารถนำไปวิเคราะห์ด้วยวิธีการทำเหมืองข้อมูลได้ เช่น ค่าเพศชายจะถูกเปลี่ยนให้เป็นค่า 1 (ชาย=1) และเพศหญิงจะถูกเปลี่ยนให้เป็นค่า 0 (หญิง=0) เป็นต้น

4.3 การสร้างแบบจำลอง (Modeling)

ในการวิจัยครั้งนี้ผู้วิจัยได้นำแบบจำลองเหมืองข้อมูลการพยากรณ์มาวิเคราะห์เพื่อจำแนกผู้ป่วยเรื้อรัง (Chronic kidney diesease) ออกจากกลุ่มอื่น โดยใช้เทคนิคเหมืองข้อมูลดังต่อไปนี้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) ป่าไม้แบบสุ่ม (Random Forest) วิธีความใกล้เคียงกันมากที่สุด (K-nearest neighbor) เทคนิคซัพ พอร์ตเวคเตอร์แมทชีน (Support Vector Machine : SVM) การเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) เครือข่ายประสาทเทียม (Neural Network) และเปรียบเทียบประสิทธิภาพของการพยากรณ์ที่ได้เพื่อหาแบบจำลองที่มีความแม่นยำในการพยากรณ์มากที่สุด ในงานวิจัยครั้งนี้ได้

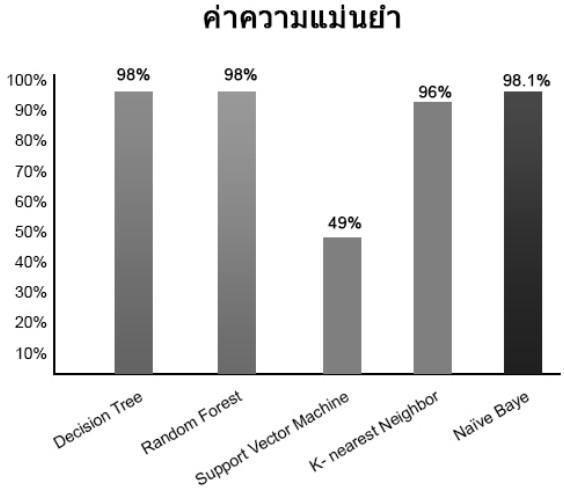
ทำการสร้างแบบจำลองโดยใช้โปรแกรม Rapid Miner และ Python(Scikit Learn) โดยข้อมูลจะถูกแบ่งแบ่งออกเป็น 2 ส่วน ในอัตตราส่วน 7 : 3 โดยใช้ส่วนที่มากกว่าในการฝึกโมเดลแบบ Supervised Learning และอีกส่วนใช้ในการทดสอบการพยากรณ์ของแบบจำลองที่ได้รับการฝึกแล้ว และทำการคัดเลือกแบบจำลองเหมืองข้อมูลที่ดีที่สุด

4.4 การประเมินประสิทธิภาพของโมเดล

การวัดประสิทธิภาพของแบบจำลองได้ใช้การประเมิณด้วย **Confusion Matrix** เป็นการประเมินผลลัพธ์การทำนาย (หรือผลลัพธ์จากแบบจำลอง) เปรียบเทียบกับผลลัพธ์จริงๆโดยคน มีค่าทั้งหมด 4 ค่า ดังนี้ True Positive (TP), True Negative (TN), False Positive (FP) และ False Negative (FN) โดยนำค่าที่ได้มาหาค่าความแม่นยำและค่า RMSE (Root Mean Square Error) ซึ่งเป็นฟังชั่นของ Scikit-learn

5. ผลและวิจารณ์

จากผลการเปรียบเทียบความแม่นยำของแต่ละแบบจำลอง พบว่า แบบจำลองนาอีฟเบย์มีค่าความแม่นยำ 98.1% และค่าความผิดพลาด 0.13736 แบบจำลองต้นไม่ตัดสินใจและป่าไม้แบบสุ่มมีค่าความแม่นยำ 98% และค่าความผิดพลาด 0.1443 แบบจำลองซัพพอร์ตเวคเตอร์แมชชีนมีค่าความแม่นยำ 49% และค่าความผิดพลาด 0.54944 แบบจำลองค่าเคใกล้เคียงมีค่าความม่านยำ 96% และค่าความผิดพลาด 0.2041 ทำให้แบบจำลองนาอีฟเบย์เป็นแบบจำลองที่มีความแม่นยำและเหมาะสมสำหรับการพยากรณ์ภาวะโรคไตเรื้อรังมากที่สุดจากแบบจำลองทั้ง 5 แบบ โดยปัจจัยสำคัญที่ทำให้ผู้ป่วยเข้าสู่ภาวะโรคไตเรื้อรัง คือ ปริมาณโซเดี่ยมในเลือดและอายุของผู้ป่วย ซึ่งคุณลักษณะหรือตัวแปรบางตัวที่มีความสำคัญต่อการจำแนกประเภทผู้ป่วยมีข้อมูลขาดหายไปส่งผลต่อความแม่นยำในการจำแนกประเภท การเพิ่มปริมาณข้อมูลเนื่องจากข้อมูลที่มี มีจำนวนไม่มากและยังต้องตัดข้อมูลบางส่วนออกทำให้ข้อมูลที่ใช้ในการฝึกแบบจำลองมีจำนวนที่น้อยซึ่งการเพิ่มข้อมูลที่ใช้ในการสร้างแบบจำลองจะสามารถเพิ่มความแม่นยำในการจำแนกผู้ป่วยได้ดียิ่งขึ้น



รูปที่ 3 แผนภูมิเปรียบเทียบความแม่นยำ

6. สรุปผล

งานวิจัยนี้ครั้งนี้มีจุดประสงค์เพื่อเปรียบเทียบความสามารถและประสิทธิภาพในการจำแนกผู้ป่วยโรคไตเรื้อรังด้วยแบบจำลอง ทำการวัดความแม่นยำและประสิทธิภาพด้วยการทำ Confuion matrix และ Classification Report และใช้เทคนิคการวิเคราะห์ข้อมูลด้วยการทำวิชวลไลเซชั่นเพื่อให้เข้าใจถึงความหมายและความสำคัญของคุณลักษณะของกลุ่มผู้ป่วย

การสร้างแบบจำลองเพื่อจำแนกผู้ป่วยโรคไตเรื้อรังด้วยแบบจำลองต้นไม้ตัดสินใจ (Dcision Tree) ป่าไม้แบบสุ่ม (Random Forest) ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine: SVM) เค – เนียเรสเนเบอร์ (K- nearest Neighbor) และ นาอีฟเบย์ (Naïve Baye) พบว่า แบบจำลอง นาอีฟเบย์ แบบ MultinomialNB มีความสามารถในการพยากรณ์ได้แม่นยำที่สุด โดยมีค่า Accuracy เท่ากับ 98.1% และค่าความคลาดเคลื่อนเท่ากับ 0.13736 ซึ่งน้อยที่สุด

จากข้อมูลผู้ป่วย 400 รายการมีจำนวนผู้ป่วยโรคไตเรื้อรังจำนวน 248 คน คิดเป็น 62% ของผู้ป่วยทั้งหมด ผู้ป่วยโรคไตเรื้อรังส่วนใหญ่อยู่ในช่วงอายุ 60-65 ปี โดยมีอัตราส่วนผู้ป่วยโรคไตเรื้อรังมากกว่า 70% ซึ่งมากกว่าผู้ป่วยในช่วงอายุอื่นๆ ส่วนกลุ่มผู้ป่วยที่มีความดันโลหิตน้อยกว่า 60 และ ความดันโลหิตมากกว่า 80 เป็นกลุ่มที่เสี่ยงที่มีโอกาศเกิดภาวะโรคไตเรื้อรังสูงซึ่งสัมพันธ์กับปริมาณโซเดียมในตัวผู้ป่วยที่มีปริมาณมาก ผู้ป่วยที่จัดอยู่ในกลุ่มที่เป็นโรคไตเรื้อรังมีปริมาณเม็ดเลือดแดงต่อปริมาณเลือดทั้งหมดโดยเฉลี่ยน้อยกว่ากลุ่มผู้ป่วยที่ไม่เป็นโรคไตเรื้อรังและยังมีค่าเฉลี่ยความดันโลหิตและปริมาณไนโตรเจนในกระแสเลือดมากกว่า

7. บรรณานุกรม

ชลนิศา สาระ. (2550). **การจำแนกกลุ่มสถานภาพการสำเร็จ การศึกษาโดยแบบจำลองต้นไม้ตัดสินใจ**. ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอม เกล้าพระนครเหนือ.

ทิพย์ธิดา วงศ์พิพันธ์.(2556).**การใช้เหมืองข้อมูลช่วยในการตัดสนิใจการให้สินเชื่อ กรณีศึกษา: บริษัท กรุงไทยคาร์เร้นท์ แอนด์ลสี จำกัด (มหาชน).** คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบณัฑิตย์

บัญชาสถิระพจน์.(2009). **Cardiovascular risks and treatment in patients with chronic kidney disease**. *Royal Thai Army Medical Journal*, 43-52.

พรรณธิภา เพชรบุญมี, ดวงกมล โพธิ์นาค และมนต์ชัย เทยีนทอง.(2556).**การพยากรณ์รูปแบบการเรียนรู้ตามประสบการณ์ของเดวดิ โคล์ป โดยใช้กฎการจำแนกเทคนิคต้นไม้ตัดสินใจ**.คณะครุศาสตร์อุตสาหกรรม มหาวทิยาลยัเทคโนโลยพีระจอม เกล้าพระนครเหนือ

อัจฉราภรณ์ จุฑาผาด.(2556).**การพัฒนาระบบสารสนเทศเพื่อการพยากรณ์จำนวนนักศึกษาใหม่ โดยใช้กฎการจำแนกต้นไม้ตัดสินใจ**. มหาวิทยาลัยราชภัฏร้อยเอ็ด

นายวุฒิชัย โนนสาคู (Woottichai Nonsakhoo) และนายปิยณัฐ ศิริสวัสดิ (Piyanat Sirisawat).(2556).**การวิจัยเชิงสำรวจในการทำนายการจราจรโดยใช้การวิเคราะห์ด้วย Big Data A Survey on Traffic Prediction based on Big Data Analytics**.ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น

โรคไตเรื้อรัง.https://www.bumrungrad.com/th/nephrology-kidney-center-bangkok-thailand/conditions/ckd-chronic-kidney-disease (14 กุมภาพันธ์ 2562)